

# Development of Digital Twins to Support the Functioning of Cyber-physical Systems

Nataliya Pankratova

Igor Golinko

## Abstract

The peculiarities of developing a digital twin for cyber-physical systems using an analytical model of the investigated process are considered. The proposed approach takes into account the conceptual uncertainty of the analytical model parameters with the subsequent passive identification by adapting the analytical model to the dynamic characteristics of the real physical process. In the article, the digital twin development is carried out on the analytical model air heating process example with the help of an electric calorifier. The analytical model uncertain parameters of the electric heater are analyzed and an integral quality index is proposed to evaluate the dynamic model adequacy in the process of air heating. For electric heater model adaptation, the passive identification algorithm of uncertain parameters is developed, in which deviations of the mathematical model uncertain coefficients are minimized. A numerical study of the considered approach has been carried out. It is shown that uncertain parameters' passive identification of the model belongs to the problem of single-extremal optimization. The considered modeling examples confirmed the effectiveness of the proposed approach to digital twins' development for cyber-physical systems.

**Keywords:** mathematical model, uncertain parameters, state space, digital twin, identification, quality criterion, cyber-physical system, electric heater.

**MSC 2020:** 68Q25.

## 1 Introduction

Cyber-physical systems (CPS) are distributed systems with a deep interconnection between their physical and computational elements. CPS can be described as intelligent systems that include computational (hardware and software) and physical components, integrated and closely interacting with each other to reflect the changing state of the real world. CPS are integrations of computation, networking, and physical processes [1]. The “brain” of the system includes billions of nodes in the form of AI and other technologies. It receives data from sensors in the real world, analyzes this data, and uses it to further control physical elements. The guaranteed functioning of CPS is based on the general problem minimization of multi-factor risks, the margin of permissible risk, forecast of the destabilizing dynamics of risk factors, principles, hypotheses, and axioms that are directly related to the analysis of abnormal situations, accidents, and disasters. The key idea of the strategy is based on the main principle: to provide timely and reliable detection and estimation of risk factors, prediction of their development during a certain period of operation, and timely identification and elimination of the causes of abnormal situations before failures and other undesirable consequences occur and prevention of the transition from normal to an abnormal mode [2]. The fundamentally important peculiarities of CPS functioning are the following: sets of risk factors and sets of situations are largely unlimited; a set of risk situations is in principle not a complete group of random events; a threshold restriction of time for decision forming is a top priority; the problem is not completely formalized; indicators of a multifactor risk estimation are not determined; criteria of a multipurpose risk minimization are not determined. The communication with computational systems and different types of sensors is implemented online in real-time. Joint actions of CPS components determine the properties and special features of the mode of functioning of a complex system at any moment of time.

To ensure the reliable operation of the CPS, a digital twin (DT) is created, which accompanies the operation of the CPS throughout its life cycle [3, 4]. When adopting a control strategy, DT allows for adequate displaying of the dynamics of the physical process, predicting

the behavior, detecting system malfunctions, finding modifications in the structure of the physical process by observable effects, and ensuring efficient and uninterrupted operation of CPS.

## 2 Relative papers

DT refers to a new innovative toolkit that helps exploit advanced scenarios of the Internet of Things (IoT) [5]. This toolkit is used to create digital copies of physical objects. These physical objects can be factories, power grids, transportation systems, buildings, cities, and more.

The number of publications on DT has increased dramatically over the past six years [6]. In total, approximately 8693 articles had been published on this topic by September 2022, but only 29 articles were published in 2016, the number had increased to 2997 by 2021 [7]. Thus, the DT development concept belongs to modern scientific trends.

One of the fundamental works on the standardization of DT development is the Industrial Internet Reference Architecture (IIRA) reference model proposed by IIC [8]. The document describes guidelines for the systems development, applications, and solutions using IoT in industry and infrastructure solutions. This architecture is abstract and provides general definitions for various stakeholders, system decomposition order, design patterns, and terms glossary. The IIRA model relates at least four stakeholder viewpoints (levels): business; usage; operation; and implementation. Each level focuses on DT functional model implementation, the structure, interfaces, and interactions between the DT components, and the DT model's system interaction with external elements of the environment to support CPS functioning. The DT technology includes (but is not limited to) combinations of the program object: physical model and data; analytical model and data; temporal variable archives; transactional data; master data; and visual models and calculations. DT creation concept has a multifaceted architecture and correspondingly complex mathematical support for implementation. Many DT-related technologies have been patented. Google Patents returned 6181 results for this query dated 2003-01-01. The largest patent holders are large corporations: Siemens AG (13.4% of all patents); General Electric (9.8%); Beijing University of Aeronautics and Astronautics (3.4%). A study that conducted a cluster

analysis of DT patents from the Webpat and Derwent databases found 140 records by 2018 [9].

In 2023, the concept of a digital twin is evolving into something more subtle and incredibly practical: an executable digital twin (xDT). Simply put, xDT is a digital twin on a chip. xDT uses data from a (relatively) small number of sensors embedded in a physical product to perform real-time simulations using reduced-order models. Based on the data from a small number of sensors, it can predict the physical state anywhere within the object (even in places where sensors cannot be placed) [10]. The world's first digital twin city was created for Singapore in 2014 at a cost of US\$73 million. In 2022, the system was replaced with an advanced version that includes data from sensors, drones, government agencies, etc. [11]. The city of Zurich has its own DT with a detailed 3D map of roads and underground and above-ground facilities [12]. According to the World Economic Forum 2022, by 2030, the technology of DT will have saved \$289 billion on city planning and construction; in 2020, investments in DT of Chinese cities exceeded \$380 billion [13].

The mathematical description of DT can be obtained using statistical modeling, machine learning, or analytical modeling techniques. Methods of statistical modeling can be divided into three groups [14]: regression analysis models; classification models; and anomaly detection models. The method choice depends on the size, quality, and data nature, as well as on the problem type and the process knowledge being modeled. For technological processes in CPS, analytical models are often used, which have valuable properties in engineering [15]. In [16], the CPS development using deterministic models, which have proven to be extremely useful, is discussed. Deterministic mathematical models of CPS are based on differential equations and include synchronous digital logic and single-threaded imperative programs. However, CPS combines these models in such a way that determinism is not preserved.

The practice of using analytical models to describe the functioning of technological processes indicates that ready-to-use models are extremely rare because such models are developed under conditions of conceptual uncertainty that need to be disclosed for a particular physical process. Conceptual uncertainty arises from the knowledge

incompleteness of the physical environment, process, or system. Conceptual uncertainty is complex [17], conceptual uncertainty examples are an uncertainties combination: objectives; operation of process; structure modeled system; system elements interaction, or interaction with the external environment, and others. For analytical models, the above uncertainties are complicated by information uncertainty, caused by methodological uncertainty (complex processes are linearized when modeling), measurement distortion (due to inaccuracy and inertia of sensors and the presence of disturbances), and other factors.

### 3 Research problem statement

Considering the above, the publication’s aim is to develop a DT model for manufacturing plants by applying the analytical model of air heating on an electric heater in conditions of conceptual uncertainty. The problem of conceptual uncertainty disclosure in contentive statement is reduced to a problem of system-coordinated disclosure of a set of diverse uncertainties on the basis of unified principles, techniques, and criteria. This set includes the uncertainty of parameters for each type of electric heater, uncertainty of its physical and mechanical characteristics, and situational uncertainty of risks in the process of operation. Such uncertainty refers to the conceptual one [17]; being distinct from information uncertainty, it represents a unified complex of the lack of information, ambiguity, and contradictoriness of interconnected and interdependent elements of a specified set of polytypic uncertainties.

## 4 Models and Methods

### 4.1 Analytical model of air heating on electric heater. Analysis of model parameters

Let’s consider the analytical model of air heating on an electric heater, which is proposed in [18]:

$$\begin{cases} T_E \frac{d\Delta\theta_E}{dt} + \Delta\theta_E = k_0 \Delta N_E + k_1 \Delta\theta_A, \\ T_A \frac{d\Delta\theta_A}{dt} + \Delta\theta_A = k_2 \Delta\theta_E + k_3 \Delta\theta_{A0} + k_4 \Delta G_A, \\ T_d \frac{d\Delta d_A}{dt} + \Delta d_A = k_5 \Delta d_{A0} + k_6 \Delta G_A; \end{cases} \quad (1)$$

here  $T_E = \frac{c_E M_E}{K_E}$ ,  $K_E = \alpha_0 F_0$ ,  $k_0 = \frac{1}{K_E}$ ,  $k_1 = 1$ ;  $T_A = \frac{c_A M_A}{K_A}$ ,  $K_A = c_A G_A + \alpha_0 F_0$ ,  $k_2 = \frac{\alpha_0 F_0}{K_A}$ ,  $k_3 = 1 - k_2$ ,  $k_4 = \frac{c_A (\theta_{A0} - \theta_A)}{K_A}$ ;  $T_d = \frac{\omega V_A}{G_A}$ ,  $k_5 = 1$ ,  $k_6 = \frac{d_{A0} - d_A}{G_A}$ .

To solve the system of differential equations (1), the state space can be used:

$$X' = AX + BU, \tag{2}$$

$$X = \begin{bmatrix} \Delta\theta_A \\ \Delta d_A \\ \Delta\theta_E \end{bmatrix}; A = \begin{bmatrix} -1/T_A & 0 & k_2/T_A \\ 0 & -1/T_d & 0 \\ k_1/T_E & 0 & -1/T_E \end{bmatrix};$$

$$B = \begin{bmatrix} k_3/T_A & 0 & k_4/T_A & 0 \\ 0 & k_5/T_d & k_6/T_d & 0 \\ 0 & 0 & 0 & k_0/T_E \end{bmatrix}; U = \begin{bmatrix} \Delta\theta_{A0} \\ \Delta d_{A0} \\ \Delta G_A \\ \Delta N_E \end{bmatrix}.$$

For models (1) and (2), the parameters classification is proposed in the block diagram form, which is shown in Figure 1. The analysis of numerical values of the model parameters allows us to conclude that thermophysical values of material flows and constructional materials of the electric heater are determined with high accuracy from handbooks on thermophysical properties of substances and materials; these parameters refer to blocks 1 – 3 (see Fig. 1).

The modifiable in general case uncertain parameters of the model are in Block 4. Formally, for model (2), there are six changing parameters  $\alpha_0$ ,  $G_A$ ,  $\theta_{A0}$ ,  $\theta_A$ ,  $d_{A0}$ , and  $d_A$ , on which all coefficients of mathematical model (2) depend. The task of numerical values finding of the parameters  $\alpha_0$ ,  $G_A$ ,  $\theta_{A0}$ ,  $\theta_A$ ,  $d_{A0}$ , and  $d_A$  should be solved in conceptual uncertainty conditions because these parameters are linked into a single complex of the mathematical model interrelated parameters.

For example, an increase in air flow  $G_A$  leads to an increase in the heat transfer coefficient  $\alpha_0$ ; exactly the same effect can be achieved by increasing the air humidity  $d_{A0}$ . The heat transfer coefficient  $\alpha_0$  depends on many factors and can significantly change its value depending on air moisture  $d_A$ ; air flow rate  $G_A$ ; temperature difference  $\theta_E - \theta_A$ ; design features of the heat exchange surface, and other factors.

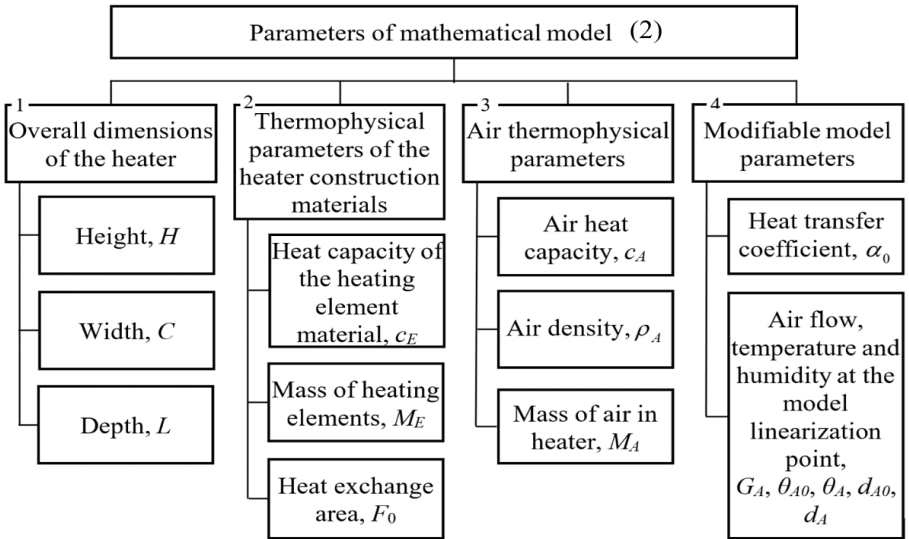


Figure 1. Block diagram of the analytical model parameters classification of electric heater

Thus, the same numerical value  $\alpha_0$  has an infinite number of combinations of the considered parameters. Note that heat transfer coefficient  $\alpha_0$  depends on many factors and there is no sensor to measure this parameter. This parameter is a research subject in thermal engineering, which is based on experimental studies and similarity theory [19]. To solve such problems, it is necessary to formulate a strategy for developing an analytical model, which will later be used to design the DT model. In these conditions, a choice of an analytical model from the available set, the uncertain parameters analysis of the model and their justification, and the method of identifying uncertain parameters is a nonformalizable procedure, and only the researcher can carry it out. The result depends on the competence, skill, experience, intuition, and other individual qualities of an actual researcher who is carrying out the given procedure.

For example, for model (2), the list of uncertain parameters contains six items, which complicates the search. However, the parameters  $G_A$ ,

$\theta_{A0}$ ,  $\theta_A$ ,  $d_{A0}$ , and  $d_A$  can be measured using CPS sensors and thus solve the uncertainty problem of these quantities. In this case, for model (2), it is necessary to reveal the uncertainty of only one parameter  $\alpha_0$ , which greatly simplifies the search task.

## 4.2 Passive identification of mathematical model parameters

Analytical model (1) is obtained using the studied laws of heat and mass transfer. For adapting the model to the concrete process of air heating, specification of its parameters is required. Identification of model parameters can be carried out using active or passive experiment on the operating equipment. Passive identification methods are most often used, since there is no need to expend additional production resources in the course of the experiment and this approach is justified by its cost-effectiveness. To identify the model in the state space (2), the Kalman filter is most often used, or the least squares method (LSM) and its modifications are used [20]. In our case we will apply LSM, taking into account its universality.

Formally, model (2) has six changing parameters  $\alpha_0$ ,  $G_A$ ,  $\theta_{A0}$ ,  $\theta_A$ ,  $d_{A0}$ , and  $d_A$  (see Fig. 2). In the passive identification process, it is sufficient to specify  $\alpha_0$  and  $G_A$ . Other uncertain parameters of the electric heater model  $\theta_{A0}$ ,  $\theta_A$ ,  $d_{A0}$ , and  $d_A$  can be estimated using the CPS sensors. As an identification criterion, using dependence that minimizes the error square of the state variables, measured values of the physical process  $X$ , and the output vector  $\bar{X}$  estimates of the model being identified (2) for the same input action  $U$ :

$$I = M \left\{ \int_{t_0}^{t_0+t_f} (X - \bar{X})^T Q (X - \bar{X}) dt \right\} \rightarrow \min, \quad (3)$$

here  $t_0$  is the initial time of the trend,  $t_f$  is the duration of the trend sampling,  $Q$  is the unit square matrix,  $T$  is the matrix transpose operator,  $M$  is the mathematical expectation operator, which takes into account industrial perturbations.

The general block diagram of passive identification of the analytical model parameters (2) is shown in Figure 2. The search criterion (3) is



calculated numerically when implementing the identification algorithm. Therefore, it is recommended to use zero-order numerical optimization methods to identify the model parameters [21]. Numerical methods require significant computational resources, so the identification algorithm can be implemented within a decision support system (DSS) [22].

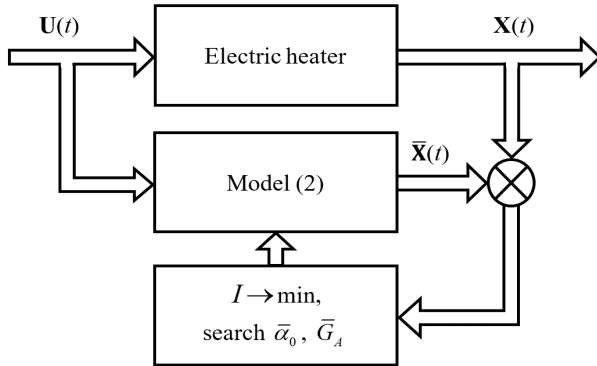


Figure 2. The block diagram of passive identification of the model parameters (2)

The passive parametric identification algorithm of the dynamic model (2) consists of the following steps:

- 1) the CPS sensors monitor the vector of input influence  $U(t)$  and output state  $X(t)$  of the physical process in real-time, the initial data are processed in the DSS; the database of input and output states for the physical process is formed;
- 2) the output state  $\bar{X}(t)$  of the mathematical model with the initial values of the parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$  is estimated by the time trends of the input action  $U(t)$ ;
- 3) the identification criterion (3) is defined, using vectors of output  $X(t)$  and the identifiable physical process states  $\bar{X}(t)$ ;
- 4) the parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$  of the identified model (2) are numerically optimized using criterion (3);

- 5) if the minimum of criterion (3) is found, then proceed to the design of DT, else continue to minimize the identification criterion and go to Step 2.

In the identification process, it is necessary to take into account the numerical method peculiarities for minimization of criterion (3). Also, for qualitative identification, the time trends  $t_f$  of input and output states should be several times longer than the physical process transient's duration. To eliminate overflow effects of the numerical minimization algorithm of criterion (3), it is necessary to set search limits for identifiable parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$ , based on the mathematical model physical feasibility.

### 4.3 Results estimation of mathematical model parameters identification

The proposed numerical identification algorithm was investigated using MatLAB. To model the time trends of the operating electric heater  $X(t)$ , the reference model (2) was used with numerical values of matrices:

$$A = \begin{bmatrix} -2.631 & 0 & 0.268 \\ 0 & -2.357 & 0 \\ 0.179 & 0 & -0.178 \end{bmatrix}; B = \begin{bmatrix} 2.362 & 0 & -21.98 & 0 \\ 0 & 2.358 & 0 & 0 \\ 0 & 0 & 0 & 0.0036 \end{bmatrix}. \quad (4)$$

To simulate production disturbances, a random signal with an amplitude of  $\pm 0.2$  was mixed into the reference output variables of the vector  $X(t)$ . In the identified model, the initial values of the parameters  $\bar{\alpha}_0 = 110$  and  $\bar{G}_A = 0.15$  differed significantly from the reference values  $\alpha_0 = 161$  and  $G_A = 0.43$ . Therefore, the numerical values of matrices  $A$  and  $B$  of the reference model (2) were significantly different from the numerical values of the matrices of the identified model:

$$\bar{A} = \begin{bmatrix} -1.265 & 0 & 0.167 \\ 0 & -1.097 & 0 \\ 0.111 & 0 & -0.111 \end{bmatrix}; \bar{B} = \begin{bmatrix} 1.099 & 0 & -21.98 & 0 \\ 0 & 1.097 & 0 & 0 \\ 0 & 0 & 0 & 0.0036 \end{bmatrix}. \quad (5)$$

For the numerical identification of  $\bar{\alpha}_0$  and  $\bar{G}_A$ , the MatLAB function `fminsearch(...)` was used, where the simplex Nelder–Mead optimization method is applied. The main results of the numerical study are presented in Figures 3 and 4.

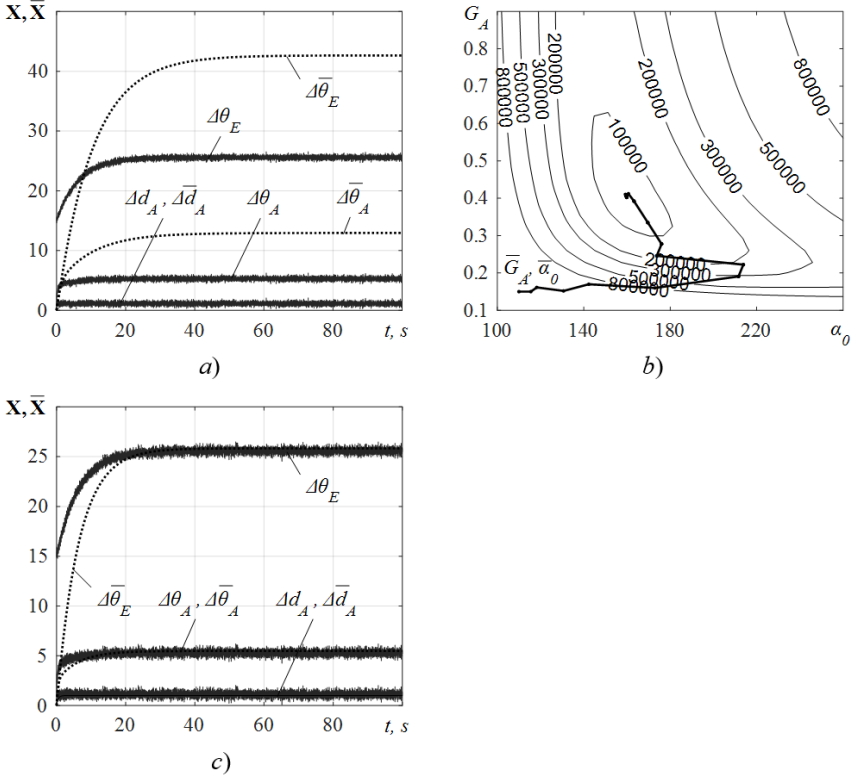


Figure 3. The parametric identification of  $\bar{\alpha}_0$  and  $\bar{G}_A$  with step influence  $U(t) = [1, 1, -0.2, 1]^T$ :  
 a) the simulation of transients before identification;  
 b) the identification trajectory of parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$  by criterion (3);  
 c) the simulation of transients after identification parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$

Figure 3 shows the case when the initial conditions of the reference model  $X(0) = [2, 1, 15, ]^T$  and the identified model  $\bar{X}(0) =$

$[0, 0, 0]^T$  are significantly different. Also, the identification parameters are significantly different:  $\alpha_0 = 161$ ,  $G_A = 0.43$ ;  $\bar{\alpha}_0 = 110$ ,  $\bar{G}_A = 0.15$ . With the step influence of the input vector  $U(t) = [1, 1, -0.2, 1]^T$ , there were obtained the transients shown in Figure 3(a). The difference between the output values of the reference  $X(t)$  and the identifiable model  $\bar{X}(t)$  is quite significant. Figure 3(b) shows the isolines surface of criterion (3) and its minimization trajectory. It is seen that the functional has one extremum in the area of the model physical practicability, so any numerical optimization method can be used as an optimization method. According to the proposed identification algorithm,  $\bar{\alpha}_0 = 160.54$  and  $\bar{G}_A = 0.4$  were determined. The found parameter values are quite close to the reference ones  $\alpha_0 = 161$  and  $G_A = 0.43$ . Figure 3(c) shows the time characteristics of the state variables of the reference  $X(t)$  and the identified  $\bar{X}(t)$  model after identification under the input influence  $U(t) = [1, 1, -0.2, 1]^T$  on both models. Based on the simulation results, it can be concluded that the proposed algorithm for passive identification of the electric heater has good convergence in the case of different initial conditions and the presence of random perturbations.

Fig. 4 (a) shows the case when the reference model is in the stationary state  $X(0) = [2.3, 0, 22.5]^T$  during the presence of random perturbations. This state is provided by the input influence vector  $U(t) = [0, 0, 0, 1]^T$ . The initial conditions of the identifiable model are zero  $\bar{X}(0) = [0, 0, 0]^T$ . According to the simulation condition, the parameters of the identified model  $\bar{\alpha}_0 = 250$ ,  $\bar{G}_A = 0.8$  are significantly different from the reference model  $\alpha_0 = 161$ ,  $G_A = 0.43$ . Fig. 4 (b) shows the surface isolines of criterion (3) and its minimization trajectory. During the identification process, the values of the parameters  $\bar{\alpha}_0 = 157.28$ ,  $\bar{G}_A = 0.42$  are optimized. As in the first study, the found values of the parameters are quite close to the reference ones. Fig. 4 (c) depicts the temporal characteristics of the state variables of the reference  $X(t)$  and the identifiable  $\bar{X}(t)$  model after identification under the input influence  $U(t) = [0, 0, 0, 1]^T$ . According to the modeling results can conclude that the proposed passive identification algorithm has good convergence in the case of the

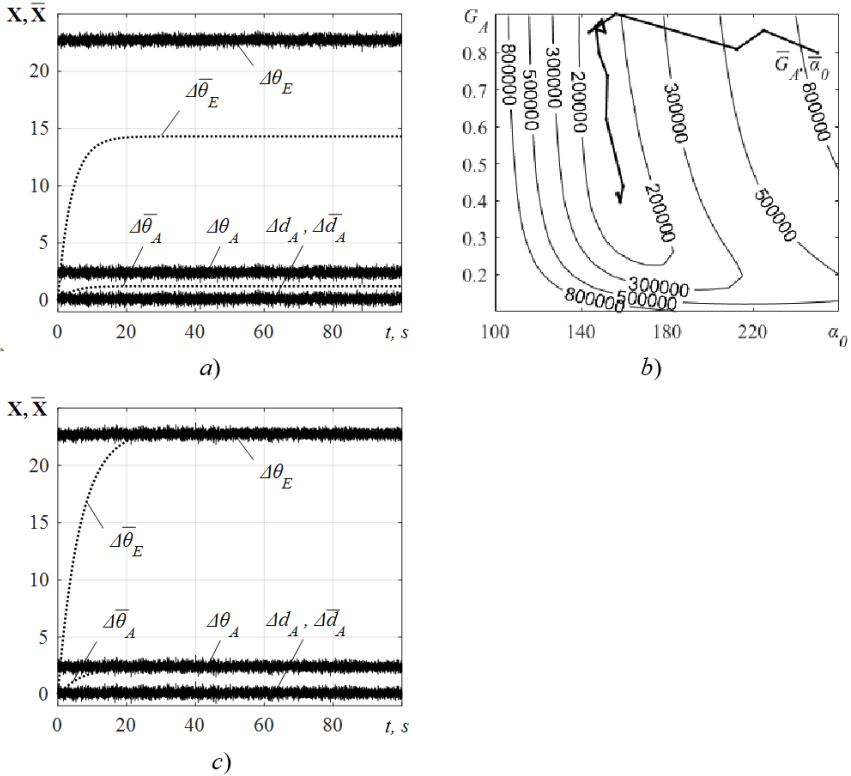


Figure 4. The parametric identification of  $\bar{\alpha}_0$  and  $\bar{G}_A$  with step influence  $U(t) = [0, 0, 0, 1]^T$ :

- a) the simulation of transients before identification;
- b) the identification trajectory of parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$  by criterion (3);
- c) the simulation of transients after identification parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$

object stationary and the random perturbations presence.

## 5 Approach to developing a digital twin of CPS under conceptual uncertainty

The Generalized structural scheme of the DT development procedure based on the analytical model of physical process is presented in Figure 5. This takes into account the parametric uncertainty of the physical process mathematical description.

In the first stage of development, it is necessary to conduct a literature analysis in the applied field of research for the physical process. This will help to determine the model structure, existing advantages, and disadvantages. As a rule, the analytical model of the studied process has the system form of differential, difference, or algebraic equations.

The practice of using analytical models shows that ready-to-use models are very rare. Even tested models require adjustment of parameters in order to adapt them to specific conditions of use. Thus, when DT is developed for a particular physical process, the researcher needs to determine the uncertainty "physical limits of that process" in the numerical values form of mathematical model parameters. To do this, the researcher needs to perform passive identification of the mathematical model parameters. A very important role at this stage is played by the data quality for the model identification, so the formation of the database should be guided by the known requirements of informativeness, synchronicity, and correctness.

The last step in DT model development is the identified model discretization. Here it is necessary to set correctly the sampling time for the mathematical model. On the one hand, the sampling time should not be small in order to ensure the information distribution over the CPS network. On the other hand, a large sampling time will lead to the loss of intermediate information for short-term forecasts. The obtained numerical model, even of a sufficiently adequate high degree, does not yet guarantee a prediction estimate of high quality if the basic uncertainties for the mathematical model of the physical

process are not taken into account. Therefore, after designing DT, it is necessary to check the possibility of using it for solving the assigned forecasting tasks.

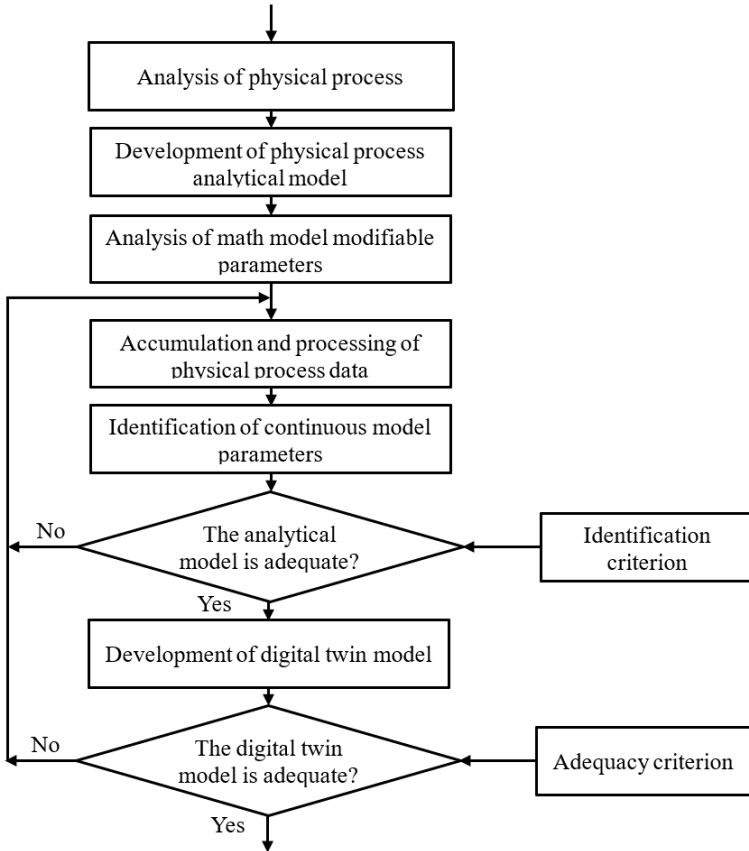


Figure 5. The structural scheme of the model DT development procedure

An important characteristic for DT is to determine the received prediction quality. Often, the quality of forecast estimates is determined with the help of LSM. However, LSM is one of many possible statistics that depends on the data scale. Therefore, only this characteristic is not enough for the analysis of a qualitative prediction. The quality

of linear and pseudo-linear models is assessed using several statistical quality criteria [20] since each criterion has its own specific purpose and characterizes one property of a prediction evaluation. Therefore, DT developer must comprehensively study a physical process, existing mathematical models, possible perturbing effects on a physical process and justify the use of adequacy criteria for DT in conditions of conceptual uncertainty.

### 5.1 Digital twin development for the air heating process by an electric heater

The DT development will be carried out on the basis of analytical model (2). In order to obtain adequate computational data at the first stage of DT synthesis, it is necessary to identify the continuous model of the electric heater, using the passive identification algorithm discussed above. Also, it is highly desirable to reduce the computational resources of DT. It is known from modeling theory that computational resources for simulation differential equations are more than for their discrete analogs based on difference equations. Therefore, let us consider a discrete representation of a continuous model (2).

The mathematical model (2) can be represented in a discrete form [23]

$$\bar{X}_{k+1} = \bar{A}_d \bar{X}_k + \bar{B}_d U_k, \quad (6)$$

here  $\bar{A}_d = e^{\bar{A}T_{KV}}$ ,  $\bar{B}_d = \int_0^{T_{KV}} e^{\bar{A}(T_{KV}-t)} \bar{B} dt$ ,  $T_{KV}$  is the sampling period.

Thus, the DT synthesis methodology for the electric heater consists of steps:

- 1) the uncertain parameters identification ( $\bar{\alpha}_0$  and  $\bar{G}_A$ ) of a mathematical model (2) by the considered algorithm;
- 2) transition from the continuous model (2) to the discrete model (6), which is DT;
- 3) if, during operation, the DT accuracy has deteriorated (due to non-stationarity of the physical process), then go to Step 1 to identify the parameters of the model.



## 5.2 Results of the digital twin simulation

The proposed methodology was used to develop and simulate the DT of an electric heater using the MatLAB software package. Let's consider the example of DT simulation using the model in state space (2) [24]. MatLAB was used to calculate the matrices  $\bar{A}_d$  and  $\bar{B}_d$  of DT (6). Simulation results are shown in Figure 6.

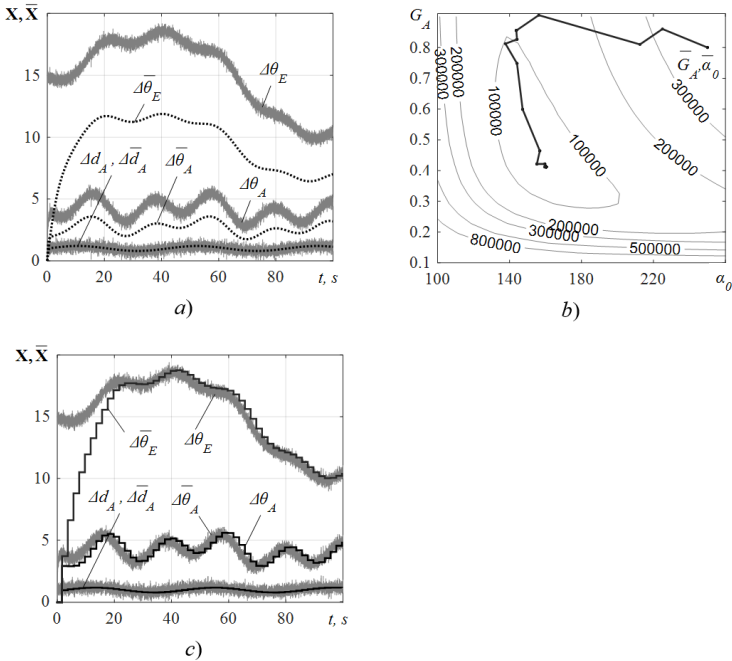


Figure 6. The DT development for the electric heater HE 36/2:  
 a) the simulation of transients before identification;  
 b) the identification trajectory of parameters  $\bar{\alpha}_0$  and  $\bar{G}_A$  according to the proposed algorithm;  
 c) the simulation of transients for the physical model (2) and DT (6)

Figure 6 (a) simulates the case with input influence

$$U(t) = [1 + 0.5 \sin(0.15t) \quad 1 + 0.2 \sin(0.15t) \quad -0.2 + 0.1 \sin(0.3t) \quad 0.5 + 0.2 \sin(0.05t)]^T,$$

and initial conditions for the physical model  $X(0) = [ 2, 1, 15, ]^T$ ,  $\alpha_0 = 161$ ,  $G_A = 0.43$  and the identifiable model  $\bar{X}_k(0) = [0, 0, 0, ]^T$ ,  $\bar{\alpha}_0 = 250$ ,  $\bar{G}_A = 0.8$ . Figure 6 (b) shows the surface isolines of criterion (6) and the minimization trajectory of its parameters  $\bar{\alpha}_0$ ,  $\bar{G}_A$ , which resulted in the finding  $\bar{\alpha}_0 = 160.3$ ,  $\bar{G}_A = 0.41$ . After identification of model (2), using MatLAB function `c2d(...)` numerical values of DT model matrices (6) are calculated for the sampling period  $T_{KV} = 2$ :

$$\bar{A}_d = \begin{bmatrix} 0.0133 & 0 & 0.0829 \\ 0 & 0.0125 & 0 \\ 0.0552 & 0 & 0.7238 \end{bmatrix};$$

$$\bar{B}_d = \begin{bmatrix} 0.9037 & 0 & -9.0374 & 0.547 \\ 0 & 0.9875 & 0 & 0 \\ 0.221 & 0 & -2.2097 & 6.1755 \end{bmatrix}.$$

Figure 6 (c) shows the time characteristics of the state variables for the reference model  $X$  and the DT  $\bar{X}_k$  of the electric heater.

## 6 Conclusions

Today there is no single approach to the processes of production plants management. Development and implementation of CPS are individual tasks for each enterprise. Systems implementation for operational management of production processes inevitably leads to the creation of multidimensional dynamic systems that are able to interact effectively in a single information management space. The solution of such problems is considered in the concept of CPS, as this technology initially assumes the symbiosis of computational and physical processes.

Digitalization of physical environments and processes implies the use of complex mathematical models with uncertain parameters, which need to be corrected in order to self-adapt to specific application conditions. As an example, the passive identification technique of the electric heater analytical model with subsequent synthesis of a digital twin for the CPS is considered. Based on system analysis methodology, the results obtained are summarized and an approach to the digital twin

development using the analytical model under conditions of conceptual uncertainties is proposed.

The peculiarity of the proposed approach is the several key parameters identification of the analytical model, which are refined in the passive identification process. The use of a physical process analytical model makes it possible to abandon the search for all its parameters. It is known that in modern methods of system analysis, the choice of structure and type of model plays an important role in further research and may require a lot of time and additional information for building an adequate model. In the proposed approach, the structure of the analytical model is known, for which only the key uncertain parameters are identified from the measured variables of the real physical process.

Examples of parameter identification for the model (2) in the state space are given. It is shown that the uncertain parameters identification of the model in the state space belongs to the problem of single-extremal optimization. The procedure for synthesizing the model of the electric heater digital twin is proposed and numerically investigated. The simulation results confirmed the effectiveness of the proposed procedure for creating a digital twin using an analytical model.

The use of DT in CPS makes it possible to identify bottlenecks in technological processes, improve product quality, and reduce the risks of abnormal operation throughout the life cycle of equipment. DT is used for the prediction of equipment operation modes and self-diagnostics, as well as optimization of the physical system structure. This approach provides a high-precision assessment of the plant's production capacity when drawing up the production program.

## References

- [1] "Cyber-Physical Systems – a Concept Map". [Online]. Available: <https://ptolemy.berkeley.edu/projects/cps/>.
- [2] N. D. Pankratova, "Creation of Physical Models for Cyber-Physical Systems," in *Cyber-Physical Systems and Control, CPS&C 2019* (Lecture Notes in Networks and Systems, vol. 95), D. Arseniev, L. Overmeyer, H. Kälviäinen, and B. Katalinić, Eds.

- Springer, Cham., 2020, [https://doi.org/10.1007/978-3-030-34983-7\\_6](https://doi.org/10.1007/978-3-030-34983-7_6).
- [3] M. Grieves, “Origins of the Digital Twin Concept,” Florida Institute of Technology, 2016. DOI: 10.13140/RG.2.2.26367.61609.
  - [4] M. Grieves and J. Vickers, “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems,” in *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, F.-J. Kahlen, S. Flumerfelt, and A. Alves, Eds. Cham: Springer International Publishing, 2017, pp. 85–113. DOI: 10.1007/978-3-319-38756-7\_4.
  - [5] N. Pankratova and Y. Perederii, “Creation of the IoT architecture for solving the problem of guaranteed functioning of the cyber-physical system,” *Information Theories and Applications*, vol. 28, no. 3, pp. 203–218, 2021.
  - [6] K. Kukushkin, Y. Ryabov, and A. Borovkov, “Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling,” *Data*, vol. 7, no. 12, 2022. DOI: 10.3390/data7120173.
  - [7] N. D. Pankratova, K. D. Grishyn, and V .E. Barilko, “Digital twins: stages of concept development, areas of use, prospects,” *System research&Information technologies*, vol. 2, pp.7–21, 2023. DOI: 10.20535/SRIT.2308-8893.2023.2.01.
  - [8] “The Industrial Internet Reference Architecture,” An Industry IoT Consortium Foundational Document, 2022. Available: <https://www.iiconsortium.org/wp-content/uploads/sites/2/2022/11/IIRA-v1.10.pdf>.
  - [9] K.-J. Wang, T.-L. Lee, and Y. Hsu, “Revolution on digital twin technology—a patent research approach,” *The International Journal of Advanced Manufacturing Technology*, vol. 107, pp. 4687–4704, 2020. DOI: 10.1007/s00170-020-05314-w.
  - [10] S. Ferguson, “Five reasons why Executable Digital Twins are set to dominate engineering in 2023,” Siemens, January 2023. [Online]. Available: <https://blogs.sw.siemens.com/simcenter/the-executable-digital-twin/>.
  - [11] Singapore government programme “Virtual Singapore”. [Online]. Available: <https://www.nrf.gov.sg/programmes/virtual-singapore>.

- [12] G. Schrotter and C. Hürzeler, “The Digital Twin of the City of Zurich for Urban Planning,” *PFG – Journal of Photogrammetry Remote Sensing and Geoinformation Science*, vol. 88, no. 2, 2020. DOI: 10.1007/s41064-020-00092-2.
- [13] World Economic Forum “Digital Twin Cities: Framework and Global Practices,” Report April 2022, Available: [https://www3.weforum.org/docs/WEF\\_Global\\_Digital\\_Twin\\_Cities\\_Framework\\_and\\_Practice\\_2022.pdf](https://www3.weforum.org/docs/WEF_Global_Digital_Twin_Cities_Framework_and_Practice_2022.pdf).
- [14] P. I. Bidyuk, O. L. Tymoshchuk, A. E. Kovalenko, and L. O. Korshevniuk, *Decision support systems and methods*, Kyiv, Igor Sikorsky Kyiv Polytechnic Institute, 2022.
- [15] E. A. Lee, “Fundamental Limits of Cyber-Physical Systems Modeling,” *ACM Transactions on Cyber-Physical Systems*, vol. 1, no. 1, 2016, pp. 1–26. DOI: 10.1145/2912149.
- [16] E. A. Lee, “The Past, Present and Future of Cyber-Physical Systems: A Focus on Models,” *Sensors*, vol. 15, no. 3, 2015, pp. 4837–4869. DOI: 10.3390/s15030483.
- [17] M. Z. Zgurovsky and N. D. Pankratova, *System analysis: Theory and Applications*, Springer, 2007, 475 p., doi.org/10.1007/978-3-540-48880-4.
- [18] N. Pankratova and I. Golinko, “Electric heater mathematical model for cyber-physical systems,” *System research and information technologies*, no. 2, pp. 7–17, 2021. DOI: 10.20535/SRIT.2308-8893.2021.2.01.
- [19] Y. M. Kornienko, Yu. Yu. Lukach, I. O. Mikulonok, V. L. Rakytskyi, and G. L. Ryabtsev, *Processes and equipment of chemical technology*, NTUU “KPI”, Kyiv, 2012.
- [20] P. I. Bidyuk, O. M. Trofymchuk, and A. V. Fedorov, “Information system of decision-making support for prediction financial and economic processes based on structural-parametric adaptation of models,” in *Research Bulletin of the National Technical University of Ukraine*, Kyiv Polytechnic Institute, no. 6, pp. 42–53, 2011.
- [21] A. E. Kononyuk, *Fundamentals of optimization theory. Unconditional optimization*, Kyiv: Education of Ukraine, 2011, 544 p. ISBN: 978-966-7599-50-8.

- [22] N. Pankratova, P. Bidyuk, and I. Golinko, “Decision support system for microclimate control at large industrial enterprises,” in *CMIS-2020: Computer Modeling and Intelligent Systems*, 2020, pp. 489–498. <https://ceur-ws.org/Vol-2608/paper37.pdf>.
- [23] V. M. Dubovoi, *Identification and modeling of technological objects and control systems*, VNTU, Vinnytsia, 2012.
- [24] N. Pankratova, I. Golinko, “Approach to Development of Digital Twin Model for Cyber-Physical System in Conditions of Conceptual Uncertainty,” in *System Analysis and Artificial Intelligence* (Studies in Computational Intelligence, vol 1107), M. Zgurovsky and N. Pankratova, Eds. Springer, Cham. pp 3–25. 2023. [https://doi.org/10.1007/978-3-031-37450-0\\_1](https://doi.org/10.1007/978-3-031-37450-0_1).

Nataliya Pankratova<sup>1</sup>, Igor Golinko<sup>2</sup>,

Received November 19, 2023

<sup>1,2</sup>Igor Sikorsky Kyiv Polytechnic Institute,  
37, Prosp. Beresteyskiy, Kyiv, 03056, Ukraine

<sup>1</sup>ORCID: <http://orcid.org/0000-0002-6372-5813>  
E-mail: [natalidmp@gmail.com](mailto:natalidmp@gmail.com)

<sup>2</sup>ORCID: <https://orcid.org/0000-0002-7640-4760>  
E-mail: [golinko.igor@111.kpi.ua](mailto:golinko.igor@111.kpi.ua)

# Multilingual Fine-Grained Named Entity Recognition

Viorica-Camelia Lupancu, Adrian Iftene

## Abstract

The “MultiCoNER II Multilingual Complex Named Entity Recognition” task<sup>1</sup> within SemEval 2023 competition focuses on identifying complex named entities (NEs), such as the titles of creative works (e.g., songs, books, movies), people with different titles (e.g., politicians, scientists, artists, athletes), different categories of products (e.g., food, drinks, clothing), and so on, in several languages. In the context of SemEval, our team, *FII\_Better*, presented an exploration of a base transformer model’s capabilities regarding the task, focused more specifically on five languages (English, Spanish, Swedish, German, and Italian). We took DistilBERT (a distilled version of BERT) and BERT (Bidirectional Encoder Representations from Transformers) as two examples of basic transformer models, using DistilBERT as a baseline and BERT as the platform to create an improved model. In this process, we managed to get fair results in the chosen languages. We have managed to get moderate results in the English track (we ranked 17th out of 34), while our results in the other tracks could be further improved in the future (overall third to last).

**MSC 2020:** 68T50.

## 1 Introduction

Named entity recognition (NER) involves identifying and classifying significant tokens (words) within a given text [1]–[3]. For instance, in news articles, identifying the names of individuals, organizations, and places is often essential. The highlighted named entities in the

following example contain valuable information and can be utilized in natural language processing (NLP) applications:

*Last month **Sky West** moved to her husband's hometown in **West Virginia**.*

such as information extraction [4],[5], question answering, text summarization, machine translation, and semantic web search, which heavily rely on NER. Named entity recognition allows for the identification of named entities such as Sky West, which is particularly useful in machine translation as it prevents erroneous word-by-word translations. It is impressive that state-of-the-art NER systems rely heavily on hand-crafted features and domain-specific knowledge [6],[7]. Over the past few decades, the scope of named entity recognition has undergone significant evolution. Initially, NER was limited to the extraction of proper nouns from news-related content, such as names of people, organizations, and locations. However, with the expansion of NLP into other domains, these traditional named entity classes proved to be insufficient. For instance, articles about science or technology require additional named entity classes beyond the original three. Additionally, it's worth noting that named entities are not limited to proper nouns. In certain fields of study, like medicine, terms such as pneumonia, common cold, or cholesterol could also be considered named entities.

The MultiCoNER II shared task [8] aims at building NER systems for 12 languages, namely English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian, and German. The task has 12 monolingual tracks and a multilingual one. The dataset contains sentences from the wiki domain, which are usually short and low-context sentences [9]. Moreover, these short sentences usually contain semantically ambiguous and complex entities, which makes the problem more difficult. Usually, retrieving knowledge related to such ambiguous concepts in any form is a definite method of understanding and disambiguating them. Thus, the ideal NER model would be capable of taking on hard samples if the option of additional context information was available. The rest of the paper is organized as follows: Section 2 briefly presents studies related to NER, either in a multi-



lingual context or not; Section 3 presents the dataset, the required pre-processing, and plausible methods for it; Section 4 resumes the results of the conducted experiments, with their interpretations, followed by Section 5 with the conclusions.

## 2 Related Work

There is a limited amount of research focused on identifying entity types beyond the conventional ones (persons, locations, organizations). Complex NEs, like chemicals, ingredients, diseases, or active substances are not straightforward nouns and pose greater challenges in terms of identification [10]. They have the ability to manifest as various linguistic constituents and have a very different surface from the traditional NEs. Their ambiguity makes it challenging to recognize them. Additionally, nowadays an increasing number of individuals are sharing information online on diverse topics, highlighting the growing significance of NER for these unconventional entities, given the data collected from social media, where people openly express their interests [11], [12]. Efforts have been made to explore the capacity of contemporary NER systems to demonstrate effective generalization across diverse genres. This attempt also found out, as expected, that a notable correlation exists between the size of the training corpus and the performance of NER systems, so by having a bigger corpus, the results may be more accurate [13]. The job of handling NEs by extracting them from the text has been done by transformers. In the last few years, new technologies have appeared, including a Google research releasing mT5, their own version of a transformer, which outperforms the previously released multilingual transformers [14]. Among those, BERT is one of the most powerful unsupervised models. A multilingual variant of it, trained in over 100 languages and enhanced with context awareness thanks to a CRF layer on top, has been leveraged before for such a task with promising results [15].

The "Multilingual Complex Named Entity Recognition (Multi-CoNER)" task<sup>2</sup> was first introduced in the context of the SemEval

---

<sup>2</sup>[https://multiconer.github.io/multiconer\\_1/](https://multiconer.github.io/multiconer_1/)

2022 competition [16]. This task was divided into 13 tracks and aimed to explore techniques for recognizing complex NEs, such as titles of creative works (movies, books, songs, etc.), products, and groups. It was conducted across 11 different languages (Bangla, German, English, Spanish, Farsi, Hindi, Korean, Dutch, Russian, Turkish, and Chinese), considering both monolingual and multi-lingual scenarios. The dataset used for this task is the MultiCoNER dataset. It contains 2.3 million samples and includes data from three domains: Wikipedia sentences, questions, and search queries, along with those 11 monolingual subsets, the multilingual and code-mixed subsets. The multilingual subset consists of randomly chosen instances from all 11 languages blended together to form a unified subset. On the other hand, the code-mixed subset holds code-mixed samples, where the tagged entities originate from one language while the remaining text inside the instance is written in a different one. The dataset defines a six-class NER tagset, as follows: **LOC** – location or physical facilities; **CW** – titles of creative works such as movies, songs, and book titles; **CORP** – corporations and businesses; **GRP** – all other groups; **PER** – people names; **PROD** – consumer products.

In last year’s MultiCoNER shared task, the two winning systems employed different strategies. [17] used a large-scale retrieval approach to gather relevant paragraphs related to the target sentence, which were concatenated and used as input to a transformer-CRF system. The aim was to build a multilingual knowledge base relying on Wikipedia. That knowledge base served the purpose of offering relevant contextual information to enhance the performance of the NER model. On the other hand, [18] employed a gazetteer-augmented BiLSTM model in conjunction with a transformer model to classify target sentences. The BiLSTM was pre-trained to generate token embeddings similar to the accompanying transformer, using sequence labels based on gazetteer matches.

In the context of the SemEval 2023 competition, we decided to focus on implementing a transfer learning approach for the BERT transformer. The concept of transfer learning involves using a pre-trained large neural network in an unsupervised manner, which next is fine-tuned for a specific task. In our case, BERT is the neural net-

work pre-trained on two tasks: masked language modeling and next-sentence prediction. Therefore, we fine-tuned this network on the NER dataset provided by the organizers. The proposed implementation uses Python programming language and is based on `Transformers` package, which is backed by the three most popular deep learning libraries – `Jax`, `PyTorch`, and `TensorFlow` – with a seamless integration between them. From `Transformers` library we made use of `BertForTokenClassification`, which is a model that has BERT as its base architecture, with a token classification head on top (a linear layer on top of the hidden-states output), allowing it to make predictions at the token level, rather than the sequence level. Named entity recognition is typically treated as a token classification problem, that’s why we chose to use it.

### 3 Dataset and Methods

Although we explored a few options, we opted for the BERT transformer model for our approach. In this section, we present statistics from the dataset, as well as the steps we went through before choosing the BERT model and using the data for training.

#### 3.1 Dataset

The dataset that we are using, MultiCoNER v2, is a large multilingual dataset (2.2 million unique instances and 26 million tokens) used for NER, that includes filtered data from public resources, Wikipedia, specifically focusing on difficult low-context sentences across 12 languages and multilingual subset. Additionally, the data underwent further post-processing to enhance its quality. A snippet with annotated entities from the dataset can be seen in Figure 1 below.

The 12 languages are part of a variety of languages with diverse typologies and writing systems, including both well-resourced languages like English and low-resourced languages like Farsi. There is a separate subset for each of the 12 languages and a multilingual subset (see Table 1), which consists of randomly collected instances from all the languages combined. From each language’s test subset, a maximum

- **English:** it was described by francis walker | OtherPER in 1866 and is known from india | HumanSettlement.
- **German:** ein vermächtnis des ottomanisches reich | HumanSettlement zerstörten sozialistische volksrepublik albanien | HumanSettlement hatte einst seine eigene medresse | Facility.
- **Spanish:** édouard herriot | Politician ou la république en personne.
- **Bangla:** স্টেশনটি প্র্যাকটিফর্ম স্কিন ডোর | OtherPROD দিয়ে সজ্জিত.
- **Farsi:** بهرام دهقانی | Artist – مری شمس و نيم | VisualWork
- **French:** un couple épatant | VisualWork réalisé par lucas belvaux | Artist sorti en 2003 | WrittenWork.
- **Hindi:** यह स्त्रियान चीन | HumanSettlement के केंद्र भाग में स्थित है।
- **Italian:** inizia la carriera in serbia | HumanSettlement nello košarkaški klub sloga kraljevo | SportsGRP per poi passare all'estero.
- **Portuguese:** em 1903 ludwig roselius | OtherPER popularizou o uso de benzeno para descafeinar | Medication/Vaccine café | Drink.
- **Swedish:** 1986 | WrittenWork bildade hon den svenska popduon roxette | MusicalGRP tillsammans med per gessle | Artist.
- **Ukrainian:** межує з египтом судан | HumanSettlement і чад | HumanSettlement.
- **Chinese:** 它也由米蓋爾·德·烏納穆諾 | Politician 引用.

Figure 1. Examples from all the languages existing in MultiCoNER II

of 35,000 instances were randomly selected, resulting in a total of 358,668 instances in the multilingual test subset. MultiCoNER II expanded on the challenges of MultiCoNER by adding fine-grained NER classes and the inclusion of noisy input. The dataset defines the following NER tagset with the 33 fine-grained classes which are listed into the 6 coarse types: **Location (LOC)** – Facility, OtherLOC, HumanSettlement, Station; **Creative Work (CW)** – VisualWork, MusicalWork, WrittenWork, ArtWork, Software; **Group (GRP)** – MusicalGRP, PublicCORP, PrivateCORP, AerospaceManufacturer, SportsGRP, CarManufacturer, ORG; **Person (PER)** – Scientist, Artist, Athlete, Politician, Cleric, SportsManager, OtherPER; **Product (PROD)** – Clothing, Vehicle, Food, Drink, OtherPROD; **Medical (MED)** – Medication/-Vaccine, MedicalProcedure, AnatomicalStructure, Symptom, Disease.

The fine-grained tagset facilitates the incorporation of various types of entities, including complex entity structures like Creative Work, as well as entities that require contextual information for disambiguation, such as Scientists and Athletes within the PER coarse-grained class.

### 3.1.1 Pre-processing

We have concatenated the training data from all of the languages into a single CONLL file. Then, to make reading and processing easier, we have converted the data into CSV format. At this step, we took note of the number of 2,671,439 total entries (tokens), spread between 67 fine-grained labels that are in the BIO scheme, which stands for Beginning-Inside-Outside. Each tag indicates whether the corresponding word is

Table 1. MultiCoNER II dataset statistics

Language	Train	Dev	Test
BN-Bangla	9,708	507	19,859
DE-German	9,785	512	20,145
EN-English	16,778	871	249,980
ES-Spanish	16,453	854	246,900
FA-Farsi	16,321	855	219,168
FR-French	16,548	857	249,786
HI-Hindi	9,632	514	18,399
IT-Italian	16,579	858	247,881
PT-Portuguese	16,469	854	229,490
SV-Swedish	16,363	856	231,190
UK-Ukrainian	16,429	851	238,296
ZH-Chinese	9,759	506	20,265
MULTI-Multilingual	170,824	8,895	358,668
Total	341,648	17,790	2,350,027

inside, outside, or at the beginning of a specific named entity. This scheme is used because named entities usually comprise more than one word. Finally, we have grouped the entries by sentence number and have used this format of the data going forward with the training. This dataset had a final size equal to 166,413 in unique instances or better-said sentences.

### 3.1.2 Preparation

Having processed our dataset, it was now time to prepare it for training. We started by having two maps ready: `labels_to_ids` which would associate each unique NE tag a unique number (having 67 total tags, we simply numbered them from 0 to 66) and `ids_to_labels` being the reverse map of the first one. Then, for each pair (*sentence*, *labels*) in the dataset, we encoded the sentence’s words using a tokenizer with a padding of 128 and converted the labels to their numeric form using our first mapping. The encoded words are then converted into tensors and each of them will be associated with the numeric labels which, similarly,

are also converted into tensors. The padding values, as well as word pieces that are not in the first part of the word after tokenization, are attributed a custom value of  $-100$ . Considering the final transformed model, we ended up using the *bert-base-uncased* tokenizer. The training set was turned into a `DataLoader` instance (from `PyTorch`), and at this point, it was ready to be used.

## 3.2 Methods

With a dataset of this size, we have run into difficulties trying to emulate the recommended baseline results with our resources, as such we opted to try out different pre-trained transformer models of small size to test which one would have the potential to be scalable within our limitations. Among the most popular and lightweight ones, we have decided to develop a model of our own based on the DistilBERT transformer. Using it as a base, we have created a baseline model for English that has been fine-tuned on the EN training data and obtained decent enough results to begin building upon it. The results of this baseline model are shown in Table 2. For the training parameters, we have used a learning rate of  $1e-2$ , a batch size of 32, several epochs of 8, and a SGD (Stochastic Gradient Descent) optimizer.

Table 2. Initial fine-tuned DistilBERT weighted results

<b>Precision</b>	<b>Recall</b>	<b>F1</b>	<b>Accuracy</b>
0.61	0.59	0.57	0.89

With this experience, we went ahead and looked into what the BERT transformer would be capable of, by comparison. We have used the *bert-base-uncased* transformer model as a start and began transfer learning, this time, using the entire collection of training data for all of the languages. We were very pleased with the initial results of the model (see Table 3). This initial run used a learning rate of  $1e-05$ , a training batch size of 4, and a validation batch size of 2, just 1 epoch and the Adam optimizer.

Further testing used the same hyperparameters, with the only difference being the number of epochs we trained the model for. Thus,

Table 3. Initial fine-tuned BERT weighted results

<b>Precision</b>	<b>Recall</b>	<b>F1</b>	<b>Accuracy</b>
0.91	0.90	0.91	0.90

the best model we managed to train in the competition is a multilingual one, trained on all 12 training subsets, using a learning rate of  $1e - 05$ , a training batch size of 4, and a validation batch size of 2, 3 epochs and the Adam optimizer. For this model, we gained a training accuracy of 0.9125 and a validation accuracy of 0.9015. The training process finished in over 2 hours. Additional research made after the competition, along with further experiments regarding the dataset and hyperparameters, as well as the improved results, can be found in the "Analysis" subsection of Section 4.

## 4 Results

### 4.1 Analysis

For the practice phase of the competition, we have submitted for each track a file that contains only the predicted tag for every token. Two scores are noteworthy, one regarding each token to its predicted tag (see Table 4) and another one regarding the predicted tag being in the correct tagset (see Table 5). We can observe that, compared to the prediction of fine-grained tags for each individual in all of the languages, the coarse-grained tagset has increased scores. This indicates that while the exact tag may not be predicted, another tag within the same tagset is successfully predicted. Analyzing the macro-averaged F1 score from Table 4, we can notice that it is below 45 for languages such as Bangla, Farsi, Hindi, Ukrainian, and Chinese, which have diverse typology and writing systems, along with a smaller number of training instances and, therefore, the lower results.

As we can notice from Table 1, for some of the languages, the number of training instances is less than 16k (as most of them have). After the evaluation phase of the competition ended, the labeled test dataset was available, so to balance the dataset, instances from the test

Table 4. Macro-averaged results of practice phase for predicted **fine-grained tagset**, using the model trained on initial dataset, with 3 epochs

Lang.	Prec.	Recall	F1
EN	67.70	62.97	64.42
BN	61.78	34.49	40.17
DE	63.26	57.41	58.78
ES	66.26	60.26	62.67
FA	37.39	25.21	28.16
FR	66.92	61.05	62.54
HI	44.75	23.62	29.28
IT	68.18	63.22	64.94
PT	68.07	58.01	61.87
SV	62.94	54.69	57.30
UK	56.85	39.63	44.75
ZH	25.72	9.59	12.95
Multi	57.49	45.85	48.99

Table 5. Macro-averaged results of practice phase for predicted **coarse-grained tagset**, using the model trained on initial dataset, with 3 epochs

Lang.	Prec.	Recall	F1
EN	79.76	77.84	78.69
BN	66.96	42.57	49.84
DE	73.39	69.45	71.03
ES	74.83	70.34	72.36
FA	55.92	36.76	42.18
FR	75.47	73.02	74.19
HI	56.58	29.34	36.94
IT	78.18	74.13	76.05
PT	74.95	68.27	71.35
SV	76.16	64.65	69.28
UK	75.00	50.83	58.28
ZH	40.54	14.64	19.58
Multi	68.98	55.99	59.98

dataset were added to the training one, and therefore, new changes appeared in training and testing files for Bangla, German, Hindi, and Chinese languages (see Table 6). A new model was trained using the balanced dataset and the same values for all hyperparameters as before. Indeed, the training accuracy increased from 0.9125 to 0.9137, but not with a noticeable impact. Next, we trained another model using the 16k dataset, but this time increasing the number of epochs, from 3 to 5. After 3 hours of training, we found the new accuracy value: 0.9342. This showed a visible impact and a new question was raised:

*What happens if we further increase the number of instances, for all the languages this time?*

Analyzing the new form of the dataset, we came to the conclusion that we could increase the number of training instances for each language to approximately 25k. This led to another form of Multi-CoNER II dataset, which can be seen in Table 7. Having the second change inside the dataset, a new training process was started. For the number of epochs, we kept the same value, i.e., 5, because we clearly



Table 6. Dataset statistics after the first change (**the 16k dataset**)

Language	Train	Dev	Test
BN-Bangla	9,708+6,5=16,208	507	19,859-6,5=13,359
DE-German	9,785+6,6=16,385	512	20,145-6,6=13,545
EN-English	16,778	871	249,980
ES-Spanish	16,453	854	246,900
FA-Farsi	16,321	855	219,168
FR-French	16,548	857	249,786
HI-Hindi	9,632+6,5=16,132	514	18,399-6,5=11,899
IT-Italian	16,579	858	247,881
PT-Portuguese	16,469	854	229,490
SV-Swedish	16,363	856	231,190
UK-Ukrainian	16,429	851	238,296
ZH-Chinese	9,759+6,6=16,359	506	20,265-6,6=13,665
Total	197,024	8,895	1,965,159

noticed an improvement in the previous model. To train this model, approximately 5 hours were needed, and the value of training accuracy increased to 0.9433.

Further, we started to increase the number of epochs from 5 to 7 and train the 25k dataset again. This time, the training process lasted almost 18 hours, finishing with an accuracy of 0.9562, the best so far. Increasing the number of epochs from 7 to 8 and the batch size from 4 to 8, the training time decreased by almost 2 hours (due to the increase in batch size), but the accuracy obtained during training did not increase significantly, having a value of 0.9578. Keeping the number of epochs the same (i.e., 8), but increasing the number of batches from 8 to 64, a new model was trained in 13 hours, but unfortunately, the accuracy during training dropped to 0.9334.

Having a better-trained model, the next step was to generate prediction files for the development (dev) dataset. After obtaining predictions for all languages, they were submitted to the right track on CodaLab. The new results for the fine-grained tagset can be seen in Table 8, and the ones for the coarse-grained tagset are displayed in

Table 7. Dataset statistics after the second change (**the 25k dataset**)

Language	Train	Dev	Test
BN-Bangla	16,208+9=25,208	507	13,359-9=4,359
DE-German	16,385+9=25,385	512	13,545-9=4,545
EN-English	16,778+9=25,778	871	249,980-9=240,980
ES-Spanish	16,453+9=25,453	854	246,900-9=237,900
FA-Farsi	16,321+9=25,321	855	219,168-9=210,168
FR-French	16,548+9=25,548	857	249,786-9=240,786
HI-Hindi	16,132+9=25,132	514	11,899-9=2,899
IT-Italian	16,579+9=25,579	858	247,881-9=238,881
PT-Portuguese	16,469+9=25,469	854	229,490-9=220,490
SV-Swedish	16,363+9=25,363	856	231,190-9=222,190
UK-Ukrainian	16,429+9=25,429	851	238,296-9=229,296
ZH-Chinese	16,359+9=25,359	506	13,665-9=4,665
Total	305,024	8,895	1,857,159

Table 8. Macro-averaged results of dev files for predicted **fine-grained tagset**, using the model trained on 25k dataset, with 7 epochs

Lang.	Prec.	Recall	F1
EN	69.55	68.20	68.26
BN	77.14	69.88	72.56
DE	73.96	71.02	71.72
ES	71.12	67.56	68.48
FA	51.64	40.62	43.55
FR	71.83	66.66	68.29
HI	72.58	57.97	63.47
IT	69.31	67.89	68.22
PT	70.09	66.39	67.57
SV	69.33	66.38	66.92
UK	67.46	52.55	57.04
ZH	38.11	16.75	22.16
Multi	66.82	59.29	61.49

Table 9. Macro-averaged results of dev files for predicted **coarse-grained tagset**, using the model trained on 25k dataset, with 7 epochs

Lang.	Prec.	Recall	F1
EN	79.19	80.35	79.68
BN	79.60	75.89	77.48
DE	81.66	80.17	80.78
ES	77.90	75.73	76.79
FA	61.24	52.35	55.79
FR	79.31	76.79	77.98
HI	74.02	60.27	65.93
IT	77.03	76.42	76.69
PT	74.89	74.53	74.67
SV	79.46	74.77	76.87
UK	75.13	63.20	67.91
ZH	53.75	24.00	31.50
Multi	74.41	67.86	70.15

Table 9. Comparing the new results with the previous ones, we can see that for Bangla and Hindi (two of the four imbalanced datasets), the macro-averaged F1 score (the score according to which the organizers evaluated and ranked the systems) obtained for the fine-grained tagset increased by more than 30 points. Next, considerable changes appeared for German, Farsi, Ukrainian, and Multilingual datasets, with an increase in macro-averaged F1 score between 12 and 15 points. For English, Spanish, French, Italian, Portuguese, Swedish, and Chinese, there were improvements, but smaller, of 3, 5, or 9 points in the macro-averaged F1 score.

In conclusion, we can confirm that increasing the training dataset really helps to improve the model and implicitly to obtain better results. Besides this, another important factor is the number of times that the learning algorithm works through the entire training dataset (the number of epochs), which in this case led to visible improvements in results.

## 4.2 Evaluation

We were able to get results for all languages in the practice phase, however, simulated errors were added in the test dataset (in 30% of the set for the following languages: English, Spanish, French, Italian, Portuguese, Swedish, and Chinese), in the evaluation phase, and our model could not handle them properly. Character-level corruption strategies were enforced for Chinese, where characters were replaced with visually or phonetically similar ones. Token-level corruption strategies, on the other hand, were devised for other languages, focusing on common typing mistakes made by humans. This involved randomly substituting a letter with a neighboring letter on the keyboard, taking into account the specific keyboard layouts of each language.

On a small scale (2-3 characters), we were able to deal with those problematic characters, but in languages that we were not familiar with, we had difficulty detecting them. Similarly to the practice phase, Tables 10 and 11 are the results we have achieved with our model during the evaluation phase for the languages where we could successfully handle the input.

Table 10. Macro-averaged results of evaluation phase for predicted **fine-grained tagset**, using the model trained on initial dataset, with 3 epochs

<b>Lang.</b>	<b>Prec.</b>	<b>Recall</b>	<b>F1</b>	<b>Ranking</b>	<b>F1 (winning team)</b>
EN	63.76	60.62	61.75	17/34	85.53
DE	57.11	55.92	55.86	13/17	88.09
ES	57.25	53.17	54.51	16/18	89.78
IT	58.85	55.99	56.36	14/15	89.79
SV	55.88	51.59	52.12	15/16	89.57

Table 11. Macro-averaged results of evaluation phase for predicted **coarse-grained tagset**, using the model trained on initial dataset, with 3 epochs

<b>Lang.</b>	<b>Prec.</b>	<b>Recall</b>	<b>F1</b>
EN	75.88	74.30	75.05
DE	72.55	69.57	70.95
ES	70.32	65.00	67.47
IT	73.19	68.01	70.39
SV	72.21	62.75	66.81

As far as rankings are concerned, in the evaluation phase we have managed to get moderate results in the English track (we ranked 17<sup>th</sup> out of 34), while our results in the other tracks could be further improved in the future: 13<sup>th</sup> out of 17 for German, 16<sup>th</sup> out of 18 for Spanish, 14<sup>th</sup> out of 15 for Italian and 15<sup>th</sup> out of 16 for Swedish. In the post-evaluation stage of the competition, we managed to achieve better results (see Table 12) with our improved system. A notable difference between the system we used during the evaluation phase and the current one is the dataset on which we trained the model. The first model was trained on the dataset that can be seen in Table 1, which contains 166,413 unique instances and 2,671,439 tokens, while the last one was trained on the 25k dataset that has 298,388 unique instances, with a total number of 7,472,249 tokens (see Table 7). With

the increase of the dataset, we also increased the number of epochs, if the first model was trained for only 3 epochs, the last one was trained for 7 epochs (the rest of the hyperparameters remained the same). The training process for the model used in the evaluation phase took over 2 hours, while it took almost 18 hours for the current one.

Table 12. Macro-averaged results after the evaluation phase for predicted **fine-grained tagset**, using the model trained on 25k dataset, with 7 epochs

Lang.	Prec.	Recall	F1	Ranking
EN	74.71	74.61	74.66	7/34
DE	75.69	74.96	75.33	7/17
ES	68.35	67.58	67.96	9/18
IT	77.96	76.64	77.30	4/15
SV	72.58	70.46	71.51	9/16

Comparing the macro-averaged F1 scores achieved for the fine-grained tagset in the post-evaluation stage (scores from Table 12) with the ones obtained during the evaluation phase (scores from Table 10), we can notice a huge improvement of almost 21 points for Italian. For German and Swedish, the score increased by over 19 points, while for English and Spanish, with approximately 13 points. With the new results obtained, that’s how we would place ourselves on the leaderboard: 7<sup>th</sup> place out of 34 for English, 7<sup>th</sup> place out of 17 for German, 9<sup>th</sup> place out of 18 for Spanish, 4<sup>th</sup> place out of 15 for Italian, and 9<sup>th</sup> place out of 16 for Swedish.

## 5 Conclusions

In this thesis, we got the opportunity to explore a transformer model’s capabilities at dealing with NLP tasks – identification of complex (fine-grained) named entities in multiple languages, in our case – and how to handle task-specific input. More specifically, we put the classic BERT model to the test and found it to live up to its reputation as a general-purpose transformer model by managing moderate results. Moreover,

our experiments showed that we could improve the performance of a model for named entity recognition using a larger training corpus. Taking into account the fact that this is not always possible, methods that integrate additional relevant knowledge (additional context information) into transformer models may overcome this insufficiency. We have learned more about the workings of the transformer model and now have a better understanding of what tackling such a task entails with regard to approaches and resource management. Therefore, we can say that a robustly optimized pre-trained approach of BERT, such as XLM-RoBERTa, which is a retrained BERT model with improved training methodology, more data and compute power, would outperform the results we achieved with BERT.

As an overall conclusion, the fine-grained level performance was inspected by the competition organizers, and it was observed that, although the coarse classes are usually easy to identify, for example, the PER class, distinguishing the fine-grained tags poses a greater challenge due to their high ambiguity [8]. In this scenario, it was observed that pre-trained transformer models often confuse entities of the Scientist class with entities from the Artist or Politician classes. This is because these models possess a higher level of pre-trained knowledge related to Artist and Politician entities compared to Scientist entities. Therefore, the problem still remains open.

As for the future directions that could improve the results of this particular model, another important thing would surely be a more versatile module for handling input test data. Contrary to expectation, we should have put more focus on this part of the system. Apart from that, parallelization of the system could have potentially made it available to us to harness more powerful transformer models. In addition to the aforementioned significance of external data, another essential element for achieving strong performance would be the usage of ensemble learning strategies: training multiple models and combining them in an ensemble to generate the final predictions.

## References

- [1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 260–270. *CoRR*, *abs/1603.01360*, no. 1603.01360. DOI: 10.18653/v1/N16-1030.
- [2] L. Zhang, X. Nie, M. Zhang, M. Gu, V. Geissen, C.J. Ritsema, D. Niu, and H. Zhang, “Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A Deep learning approach,” *Frontiers in Plant Science*, vol. 13, 2022, Article ID: 1053449. DOI: <https://doi.org/10.3389/fpls.2022.1053449>.
- [3] A. Iftene, D. Trandabăț, M. Toader, and M. Corîci, “Named Entity Recognition for Romanian,” in *Proceedings of the 3th Conference on Knowledge Engineering: Principles and Techniques Conference (KEPT2011)*, Studia Universitatis, Babes Bolyai, vol. 2, 2011, pp. 19–24.
- [4] Y. Shao, C. Hardmeier, and J. Nivre, “Multilingual named entity recognition using hybrid neural networks,” in *The Sixth Swedish Language Technology Conference (SLTC)*, 2016, <https://api.semanticscholar.org/CorpusID:57588814>.
- [5] D. Gifu and G. Vasilache, “A language-independent named entity recognition system,” in *Proceedings of The 10th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2014*, Alexandru Ioan Cuza” University Publishing House, Iași, 2014, pp. 181–188.
- [6] D. Cristea, D. Gifu, I. Pistol, D. Sfirnaciuc, and M. Niculiță, “A mixed approach in recognising geographical entities in texts,” in *Linguistic Linked Open Data: 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop*, (Sibiu, Romania, July 13–25, 2015), Revised Selected Papers 1, Springer, 2016, pp. 49–63.

- [7] A. Iftene, “Identifying Geographical Entities in Users’ Queries,” in *CLEF 2009, LNCS 6241, Part I (Multilingual Information Access Evaluation, Vol. I, Text Retrieval Experiments)*, C. Peters et al., Eds. Heidelberg: Springer, 2010, pp. 535–538.
- [8] B. Fetahu, S. Kar, Z. Chen, O. Rokhlenko, and S. Malmasi, “SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2),” in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, ACL, 2023.
- [9] B. Fetahu, Z. Chen, S. Kar, O. Rokhlenko, and S. Malmasi, “MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition,” arXiv:2208.14536, 2023.
- [10] M. Mitrofan and V. Pais, “Improving Romanian BioNER Using a Biologically Inspired System,” in *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, 2022, pp. 316–322.
- [11] S. Ashwini and J. D. Choi, “Targetable Named Entity Recognition in Social Media,” arXiv:1408.0782, 2014.
- [12] A. Iftene and A. Balahur-Dobrescu, “Named Entity Relation Mining Using Wikipedia,” in *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, (28-30 May, Marrakech, Morocco), 2008, [http://www.lrec-conf.org/proceedings/lrec2008/pdf/192\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/192_paper.pdf).
- [13] I. Augenstein, L. Derczynski, and K. Bontcheva, “Generalisation in named entity recognition: A quantitative analysis,” *Computer Speech & Language*, vol. 44, pp. 61–83, 2017.
- [14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mT5: A massively multilingual pre-trained text-to-text transformer,” arXiv:2010.11934, 2020.
- [15] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, “Tuning Multilingual Transformers for Language-Specific Named Entity



- Recognition,” in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, 2019, pp. 89–93.
- [16] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko, “SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER),” in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Association for Computational Linguistics, 2022, pp. 1412–1437.
- [17] X. Wang, Y. Shen, J. Cai, T. Wang, X. Wang, P. Xie, F. Huang, W. Lu, Y. Zhuang, and K. Tu, “Damo-nlp at Semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition,” arXiv:2203.00545, 2022.
- [18] B. Chen, Y.Y. Ma, J. Qi, W. Guo, Z.H. Ling, and Q. Liu, “USTC-NELSLIP at SemEval-2022 task 11: gazetteer-adapted integration network for multilingual complex named entity recognition,” arXiv:2203.03216, 2022.

Viorica-Camelia Lupancu, Adrian Iftene

Received December 08, 2023

Viorica-Camelia Lupancu

”Alexandru Ioan Cuza” University of Iasi, Romania, Faculty of Computer Science  
General Berthelot, No. 16, Iasi, Romania  
E-mail: lupancu\_camelia\_99v@yahoo.com

Adrian Iftene

ORCID: <https://orcid.org/0000-0003-3564-8440>

”Alexandru Ioan Cuza” University of Iasi, Romania, Faculty of Computer Science  
General Berthelot, No. 16, Iasi, Romania  
E-mail: adiftene@info.uaic.ro

# Distinctive features of recognition for documents printed in the Romanian transitional alphabets

Tudor Bumbu, Lyudmila Burtseva,  
Svetlana Cojocar, u,  
Alexandru Colesnicov, Ludmila Malahov

## Abstract

In this paper, we summarize the research of digitization of documents printed by Romanian transitional alphabet. These printings are the most original Romanian historical documents, which makes our experience useful when researching OCR methods for similar alphabets.

The current work is focused to OCR that is the first stage of scanned documents digitization. The technique of OCR of documents, printed in the Romanian transitional alphabet, is presented. In particular, this technique is embedded in our digitization platform HeDy.

A series of examples is presented to demonstrate the application of the described technique.

**Keywords:** cultural heritage, OCR, Romanian transitional alphabets

**MSC 2020:** 68T50.

## 1 Introduction

The presented work concerns the modern research in European cultural heritage preservation: saving the originality of every country. The author of high detailed work *History of Romanian Spelling* [1] termed the transitional Romanian alphabet as the most original graphic systems in the modern history of European cultures.

Despite of mentioned high originality, the necessity of transitional alphabet sprang from the common European phenomenon. The majority of the first printed books were the cult books. The originality is based on a combination of two factors specific to the Romanian language. The cult books in Romanian regions are written by Cyrillic scripts. But Romanian language belongs to Latin Group and its natural script is Latin. The transition of all Romanian writing to Latin scripts was declared at 1830. But transition process was so complicated and intricate that the author of specific work [2], that concerned transitional alphabet, called it *saga*. This process was not even straightforward, being performed in *trial and error* way. After unsuccessful attempts, some printing houses temporarily returned to Cyrillic script.

Our initial research on the topic is described in [3]. The transition was performed by direct transliteration. This technique succeeded, for example, in transition of specific Cyrillic letters of diphthong:  $\text{ѣ} \rightarrow \text{ia}$ ,  $\text{ѥ} \rightarrow \text{ie}$ . But direct transliteration was rejected as representation of *phonetic* sense. In the traditional Latin orthography the *phonetic* sense is usually represented by several letters, like, for example, in English:  $\text{sh} \rightarrow /ʃ/$ . But multi-letters printing was often not accepted by readers, so direct transliteration technique was rejected. The second technique was to establish the closest possible correspondence between the sound value of the old and new letters. This technique was accepted as the main one. The transition rules have been discussed and tested for a long time. For some sounds, the multi-letter representation was replaced by the design of own, specific Romanian, letters, and as a result, the Romanian diacritics was born.

Such a complex and lengthy transition process leads to difficulties with the formal definition of the transitional alphabet. But the general definition is necessary for reference and can be formulated as follows.

**Romanian transitional alphabet** is the alphabet, that was used in Romania in 1830-1870 and was designed for transition from Cyrillic script to Latin script, defined as 36 letters, 27 of which are modern ones, plus 9 old letters: Ъ ѡ, ІѢ ѣ, А а, Ѡ ѡ, Ѣ ѣ, Ѥ ѥ, Ї ї, Ъ ѡ, ІІІ ііі, ІІ и.

Digitization of texts printed in transitional alphabets is an important element of digitization of Romanian historical printed publications. The period of using transitional alphabets coincided with the era when printed publications became a necessary part of everyday life. Thus, this period gave a rich legacy of both periodicals and literature, which can be called the first purely

Romanian. So, digitized copies of transitional alphabets prints are interesting not only for linguists but also for historic and literary researchers.

Digitization of documents printed in the Romanian transitional alphabet has specific aspects, both common to all very original documents and specific to Romanian printings.

In this paper, we summarize our experience in digitizing historical documents printed in the Romanian transitional alphabet.

Digitization was implemented by our platform HeDy [4].

## 2 OCR specific aspects

As mentioned in the preceding section, the formal definition of the transitional alphabet serves as a reference point. This is due to the fact that Cyrillic and Latin letters were mixed in varying proportions, influenced by factors such as the time period, location, and the preferences of typographers, editors, or authors of texts. Book [2] counts up to 17 variants of the transitional alphabets, while some authors declared approx. 20 variants.

Such an irregular variety of transitional alphabets creates problems for all digitization elements, but especially for OCR.

The main problem of the transitional alphabet OCR is the setting of OCR engine. To get acceptable accuracy, OCR tools have to be prepared for a particular variant of transitional alphabet, which means to be: (1) configured, (2) trained and (3) supplied by the proper dictionary.

During the development of HeDy, two approaches to transitional alphabet OCR were tested.

The firstly tested approach is to reproduce the resulted text in its original variant of transitional alphabet. AFR, prepared as it was described above, produces 7% of erroneous words. This is a good result, but the preparation process takes a lot of time and resources.

To achieve more effectiveness, the second approach was tested. This approach uses the general feature of large OCR systems, in our case AFR, to output the result both using original glyphs and substituting them by any sequence of letters from the selected alphabet of recognition. This is called ligatures in AFR documentation. So, the second approach consists in using as ligature the Latinized version of the transitional alphabet that we

specifically developed. For example, both **т** (Cyrillic) and **t** (Latin) will be recognized as **t**.

The OCR using this intermediate alphabet can be the final step if the goal is to obtain a source for transliteration. To solve problems whose purpose is to reproduce the original, we have added a utility to our platform that converts the intermediate OCR output into the desired variant of the transitional alphabet.

The second approach proved fruitful. The OCR errors reduced to 4.8%.

This approach also reduces the volume of the dictionary. For example, **trekut** (modern Latin script **trecut**) in the recognition dictionary may check up to 16 variants obtaining by independently replacing **t** → **т**, **r** → **p**, **k** → **κ**, **u** → **γ**.

The second approach as well solves technical problem of OCR engine setting. AFR, for example, does not support arbitrary Unicode glyphs in its dialogs and forms. Old Romanian letter **ⵀ** was introduced in Unicode only after 2009. Standard system fonts do not contain some Romanian Cyrillic letters. As a result, we see in AFR empty boxes instead of letters during training, alphabet formation, etc. The use of ligatures allows to employ fonts only at the stage of converting the output data, when we control the view.

### 3 OCR of Romanian transitional alphabet by examples

In this section the transitional alphabet OCR by HeDy is demonstrated by examples. The list of examples, which have some particular and interesting features, was extracted from the book [1]. The scanned sources according this list were obtained from free web bases, the links are referred at footnotes.

#### 3.1 Initial usage

The arrival of a transitional alphabet, both in Muntenia and Moldova, dates back to 1829-1830. One of the most known examples of the initial transitional alphabet usage is Iasi newspaper *Albina româneasca*. Initially, the

appearance of Latin letters was very rare, as we can see from the example<sup>1</sup> in Fig. 1, which is a fragment of the first issue on June 1, 1829. The text is practically entirely printed in Cyrillic script, except for a single Latin letter *i*. The low quality of the original text has been improved by image pre-processing; however, two recognition errors remain.



Ешіи 1 юніе 1829.

АЛБИНА РОМЪНЪСКЪТЪ  
ГАЗЕТЪ ПОЛИТИКО-ЛИТЕРАЛЪТЪ.

ДНАИНТЕ КЪВЖНТАРЕ.

Епоха д карѣ трѣим поартѣ семне  
жѣштите ши вредниче де мираре!  
Дорѣл лвъцѣтѣрилор нѣнѣмай къ  
лфрѣцѣще пе лѣкѣбитОрїи оуinei цери  
лтрѣ кжшигарѣ ачестеи моралниче  
авѣци, прин карѣ w нацїе се фаче  
пѣтертикъ ши феричитѣ, чи лѣкѣ ши  
шамени не асемънаци кѣ Релїгїа, кѣ  
лимба ши кѣ лециле сжнт

Вѣцїндѣсе ши сѣпѣндѣсе дрептелор  
лецї. Оаре пѣтемѣ нои приви ла  
ачѣсте бѣне оурмате лнаинтѣ шкилор  
ностри, фѣрѣ а ни лѣчриста къ нѣмай  
нацїа ноастрѣ лѣчѣ маи маре парте  
есте лифитѣ де ачести лѣбнѣтѣцири  
ши лнапоетѣ де кжт тоате пѣмѣрїле  
Еуропей, ши де кжт мѣлте алтеле че  
лѣкѣдескѣ пре челе лалте пѣрцї але  
пѣмжнтѣлѣй? Чине нѣ сжмте д цара  
ноастрѣ лиѣа ашезѣмжнтѣрилор

Figure 1. The first issue of *Albina româneasca*, June 1, 1829.

In later issues, especially in the literary supplement *Alăuta Româneasca*, Latin characters become more and more frequent (Fig. 2). One can see that the title is printed entirely in Latin letters and the following text is in Cyrillic with full replacement of the Cyrillic *и* with the Latin *i*. For the spelling of the letter *t*, two variants are used: *т* and *ш*. The word *кѣрѣ* is recognized erroneously as *коре*.

<sup>1</sup>The foolowing examples are taken from <https://tiparituriromanesti.wordpress.com/>



ALĂUTA ROMÂNESCĂ.  
 SUPPLEMENT LITTERAL  
 ALBINEI ROMÂNESCI.  
 IASSI. 1. IULIE 1838.

Ачест Суплемент а Газетеи, есь де доуѣ ори пе лунѣ ла

бантора Албінеі Ромънещї •Ѧн Еши.

ДѦн партеа Редакціѣ.

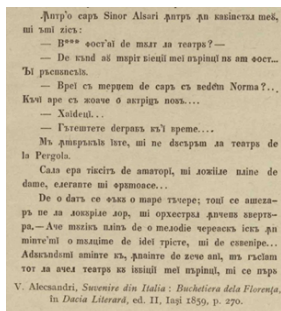
Редакція Албінеі Ромънещї, нѣ аѣ крѣчат, •Ѧн коре де ноѣ анї, нїчї о жѣртѣѣ спре а 'ші мѣлцѣмі аБОНАЦІЇ. Дорѣнд сѣ лі дее о ноѣ довадѣ де вѣна еї воїнцѣ, ел аѣ лѣат мѣѣдрїле кѣвїнчоасе ка Алѣѣгта Ромънеаскѣ, каре пѣнѣ акѣм се да неѣхотѣртї, сѣ еасѣ •Ѧн вїгорїе, регѣлат де доѣѣ ори пе лѣнѣѣ, ла 1 шї ла 15 а лѣнеї, кѣспрїнзїнд нѣмаї лѣкрѣрї лїтерале прекѣѣм ачест •Ѧнгѣї нѣмѣр. Редакція се ва сїргѣї ка Алѣѣгта сѣѣ рѣѣне прѣдѣктѣрїле дѣѣѣлѣї челе маї несѣ шї челе маї интересанте пентрѣ четїторї. Еа •Ѧшї ва •Ѧмплїнї скопосѣл кѣ аѣѣторѣл мѣлтор тїнерї лїтерарї, карїї аѣ БИНЕВоїт а фагѣдѣї лѣкрѣрїле лор ачестеї пѣблїкації перїодїче.

Figure 2. *Alăuta Româneasca*, July 1, 1838.

### 3.2 Complete usage

The next example presents the full version of the transitional alphabet how Heliade Radulescu, Balcescu, and Treboniu proposed it, and how Laurian, Asachi, Kogalniceanu, etc. used it. We selected as an example an excerpt from the famous author Vasile Alecsandri's novel *Suvenire din Italia: Buchetiera de la Florența* (*Souvenirs from Italy: A Florist from Florence*) (Fig. 3) because the works of the classical authors were re-published by modern Romanian alphabet. Modern re-publications are a useful source for verifying the transliteration tool. The transliteration tool, in turn, can be used for creation of datasets for both training and validation of the OCR engine. The quality of OCR is very good, with only one error in this fragment (the Latin letter **d** recognized as the Cyrillic letter **б**). Most of the letters are Cyrillic, with the letters **D**, **d**, **i**, **m**, **n** being written in Latin script. In the case of the last letter, the Cyrillic spelling **н** is also found. Note that all proper names: **Sinor Alsari**, **Norma**, and **Pergola** are written in Latin characters.

Another example of using the transitional alphabet in its mature state



«Пѣтр'о сарѣ Sinog AIsari Пѣтрѣ Пн кабинетѣл меѢ,  
 ши ѣмї Ѣїсѣ:  
 — В\*\*\* фост'аї де мѣлт ла театрѢ? —  
 — Де кѣнд аѢ мѢрїт бїеції меї пѣрїнци нѢ ам ФОСТ...  
 Тї рѣспѣнѣїѢ.  
 — Вреї сѣ мерѣем де сарѣ сѣ vedem Norma?... Кѣчи  
 вре сѣ жоаче о актрїцѣ поѢ...  
 — Хаїдеці...  
 — Гѣтешете деграбѣ кѣ време...  
 Мѣ ПѣбрѣкѣїѢ їѢте, ши не дѢсѣрѣм ла театрѢ де In  
 Pergola.  
 Сала ера тїкїсїѣ де аматорї, ши ложїїле плїне de dame,  
 елѣганте ши фѢрѢмоасе...  
 Де о датѣ се Ѣзак о шарѣ тѣчере; тоцї се ашеѢсарѣ  
 пе ла локѢрїле лор, ши орѢкѢстрѣл дїчѣвѣ аверѢрѣ-  
 рѣ. — Ave muzica talia de o melodie cereasca icuș din  
 mīre'mi o maxime de idei trīste, mi de esenție...  
 Adărundu-mi aminte că, fuziune de acei ani, eu găsisem  
 tot la ovel teatrul ca izvoiașii mei țipăuți, mi se țipău  
 V. Alexandri, *Souvenirs din Italia: Buchetiera din Florența*,  
 în *Dacia Literară*, ed. II, Iași 1859, p. 270.

Figure 3. Fragment of *Souvenirs from Italy: A Florist from Florence* by V. Alexandri, 1840

is interesting by the confident use of the transitional alphabet in secular literature. The example is from the novel *Radu VII from Afumați*<sup>2</sup> by Stefan Andronic (Fig. 4A, p. 347). The *Dictionary of Romanian literature* declares this novel as the first Romanian historical roman.

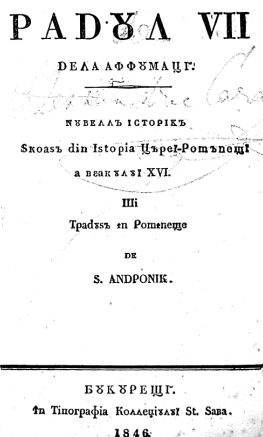
### 3.3 One direction conversion Cyrillic – Transitional – Latin

The book [1] presents the examples which are named *effective transition*. The definition of *effective* here means that the mentioned printed publications: started issuing using the transitional alphabet immediately after the announcement; used the transitional alphabet only for a few years; finally, were issued entirely in the Latin alphabet without any returns.

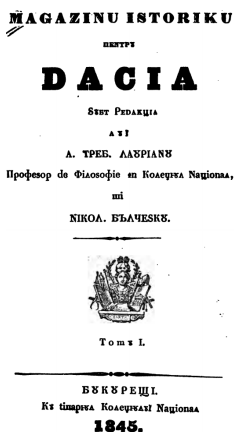
The presented example is the fragment from *Magazinu istoricu pentru*

<sup>2</sup><https://revistatransilvania.ro/wp-content/uploads/2019/11/1846.-S.-Andronic-Radu-VII-de-la-Afumatii.pdf>





A.



B.

Figure 4. A. Cover of *Radu VII from Afumați* by S. Andronic, 1846  
 B. Title page of *History Magazin for Dacia*, 1845

*Dacia*<sup>3</sup> (Fig. 4B, p. 347), edited by Laurian and Balcescu, that used the transitional alphabet in 1845–1847 and then the pure Romanian Latin alphabet.

## 4 Accuracy evaluation

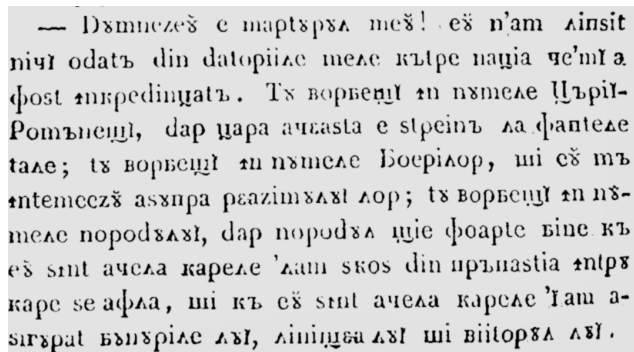
Here we propose analysis of accuracy in recognition of p. 20 of the novel *Radu VII from Afumați* (see Sec. 3.2 above).

The most frequent errors were observed in the letter **б** (changed to **в**) and the letter **н** (changed to **п**). With total 1172 characters on the page and 93 erroneous characters, we have the 92.1% accuracy.

This result could be improved by adding a user dictionary as part of the FineReader language model. With the dictionary, we got only 31 erroneous characters, and the accuracy became 97.4% that seems to be a good result.

However, after adding the dictionary, while most errors with the letter **н** were resolved (as similar words were added to the dictionary), there were still recognition errors with the letters **и**, **б**, **м**, and some others. Also, the

<sup>3</sup>[https://archive.org/download/magazinuiistoric01unkngoog/magazinuiistoric01unkngoog\\_tif.zip](https://archive.org/download/magazinuiistoric01unkngoog/magazinuiistoric01unkngoog_tif.zip)



— Дѣмнезеѣ е мартѣрѣл меѣ ! еѣ н'ам лѣнсѣт  
нѣчѣ одатѣ дѣн даторѣиле меле кѣтрѣ наѣѣа че'мѣ а  
ѣост ѣнкредѣнѣатѣ . Тѣ ворбѣнѣѣ ѣн нѣмеле Цѣрѣѣ-  
Ромѣнеѣѣѣ , дар ѣара ачеаста е стрѣнѣѣ ла ѣаптеле  
тале ; тѣ ворбѣнѣѣ ѣн нѣмеле Боерѣлор , шѣ еѣ мѣ  
ѣнтемеезѣ асѣпра реазѣмѣлѣѣ лор ; тѣ ворбѣнѣѣ ѣн нѣ-  
меле норолѣлѣѣ , дар норолѣл ѣѣѣ ѣоарте бѣне кѣ  
еѣ снѣт ачела кареле 'лам скос дѣн прѣпастѣа ѣнтрѣ  
каре се аѣла , шѣ кѣ еѣ снѣт ачела кареле 'ѣам а-  
сѣѣрат бѣнѣрѣле лѣѣ , лѣнѣѣѣа лѣѣ шѣ вѣѣторѣл лѣѣ .

— Дѣмнезеѣ е мартѣрѣл меѣ ! еѣ нам лѣнсѣт нѣчѣ одатѣ дѣн даторѣиле меле кѣтрѣ наѣѣа чемѣ а ѣост ѣнкредѣнѣатѣ . Тѣ ворбѣнѣѣ ѣн нѣмеле Цѣрѣѣ-Ромѣнеѣѣѣ , дар ѣара ачеаста е стрѣнѣѣ ла ѣаптеле тале , тѣ ворбѣнѣѣ ѣн нѣмеле Боерѣлор , шѣ еѣ мѣ ѣнтемеезѣ асѣпра реазѣмѣлѣѣ лор ; тѣ ворбѣнѣѣ ѣн нѣмеле норолѣлѣѣ , дар норолѣл ѣѣѣ ѣоарте бѣне кѣ еѣ снѣт ачела кареле лам скос дѣн прѣпастѣа ѣнтрѣ каре се аѣла , шѣ кѣ еѣ ачела кареле 'ѣам асѣѣрат бѣнѣрѣле лѣѣ , лѣнѣѣѣа лѣѣ шѣ вѣѣторѣл лѣѣ .

Figure 5. Digitization of *Radu VII from Afumati* fragment

word **ѣубенеѣ** was incorrectly recognized this second time, even though the model recognized it correctly without the dictionary.

It's worth noting that the model almost flawlessly recognizes special characters and punctuation, often erring in the letters **н**, **п**, **ш**, **м**, **б**, and **ц**. The model learned to recognize the letter **ѣ** well and also correctly recognizes the letter **л**.

The letter **ѣ** was not on this particular page, but it was often identified as **e** in other pages. In reality, later the **e** was used instead of this letter.

Later it was observed that the letter **ѣ** also appears in texts but was always replaced with **i**. The letter which looks like **ѣ** with a line over it was counted as a simple **ѣ**.

Ної кредем къ о асѣминѣ Історіе есте кѣ пѣтин-  
цѣ. Пентрѣ ачеаста сокотим къ чеї че се окѣпѣ кѣ  
Історіа нѣстрѣ, нѣ трѣвѣ а се цинѣ нѣмаї де чеа че  
аѣ лѣкрат ши аѣ зис историчїї нострїї чеї модернї;  
дар, тот нѣтр'о време, фолосиндѣсе де адеврѣрѣ дес-  
коперите де дѣншиї сѣ мѣргѣ маї департе, сѣ алер-  
це ла исвѣрѣле ориѣинале, сѣ каѣте ши сѣ адѣне тѣте  
датѣриле пѣтинчиѣсе, ши атѣнчї вор пѣтѣ цесе о Бѣнѣ  
Історіе.

Ної кредем къ о асѣминѣ Історіе есте кѣ пѣтинцѣ. Пентрѣ ачеаста сокотим къ чеї че се окѣпѣ кѣ Історіа нѣстрѣ, нѣ трѣвѣ а се цинѣ нѣмаї де чеа че аѣ лѣкрат ши аѣ зис историчїї нострїї чеї модернї; дар, тот нѣтр'о време, фолосиндѣсе де адеврѣрѣ дескоперите де дѣншиї сѣ мѣргѣ маї департе, сѣ алерце ла исвѣрѣле ориѣинале, сѣ каѣте ши сѣ адѣне тѣте датѣриле пѣтинчиѣсе, ши атѣнчї вор пѣтѣ цесе о Бѣнѣ Історіе.

Figure 6. Digitization of *Magazinu istoricu pentru Dacia* fragment

## 5 Conclusion

Digitization of historical documents, printed by rare and unique alphabets, is an important part of the preservation of a specific national cultural heritage. The solutions of OCR problems, which arise during digitizing such documents, has scientific and practical value.

These solutions are taken into attention in our HeDy platform, which exists in free versions for both web and desktop. In addition to OCR tools presented in current work, HeDy platform provides the tool for transliteration, that simplifies reading the historical documents content.

**Acknowledgments.** This work was prepared as part of the research project 20.80009.5007.22 *Intelligent information systems for solving ill-structured problems, processing knowledge and big data.*

## References

- [1] Pârvu Boerescu, *Din istoria scrierii românești*, București: Editura Academiei Române, 2014, 400 p. ISBN: 978-973-27-2459-0. (in Roma-

nian).

- [2] Ștefan Cazimir, *Alfabetul de tranziție; Jurnal de tranziție*, Oscar Print, 1996. 197 pp. ISBN: 9789739757348. (in Romanian).
- [3] S. Cojocaru, L. Malahov, A. Colesnicov, T. Bumbu, “Optical Character Recognition Applied to Romanian Printed Texts of the 18th–20th Century,” *Computer Science Journal of Moldova*, vol. 24, no. 1(70), pp.106–117, 2016.
- [4] T. Bumbu, L. Burțeva, S. Cojocaru, A. Colesnicov, L. Malahov, “A platform for processing heterogeneous documents,” in *Proceedings of the 17th International Conference “Linguistic Resources and Tools for Natural Language Processing”*, Univ. “A.I.Cuza”, Iași, 2022, pp. 141–151.

Tudor Bumbu<sup>1,2</sup>, Lyudmila Burtseva<sup>1,3</sup>,  
Svetlana Cojocaru<sup>1,4</sup>,  
Alexandru Colesnicov<sup>1,5</sup>, Ludmila Malahov<sup>1,6</sup>,

Received November 29, 2023

<sup>1</sup>“V. Andrunachievici” Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova

<sup>2</sup>ORCID: <https://orcid.org/0000-0001-5311-4464>

E-mail: [bumbutudor1@gmail.com](mailto:bumbutudor1@gmail.com)

<sup>3</sup>ORCID: <https://orcid.org/0000-0002-9064-2538>

E-mail: [luburtseva@gmail.com](mailto:luburtseva@gmail.com)

<sup>4</sup>ORCID: <https://orcid.org/0009-0003-1025-5306>

E-mail: [svetlana.cojocaru@math.md](mailto:svetlana.cojocaru@math.md)

<sup>5</sup>ORCID: <https://orcid.org/0000-0002-4383-3753>

E-mail: [acolesnicov@gmx.com](mailto:acolesnicov@gmx.com)

<sup>6</sup>ORCID: <https://orcid.org/0000-0001-9846-0299>

E-mail: [ludmila.malahov@math.md](mailto:ludmila.malahov@math.md)

# Supplementing elearning systems with adaptive content generation elements

Alexandr Parahonco, Mircea Petic

## Abstract

The paper describes automatic summarization as one of the topic that helps elearning system to be more adaptable on content generation. This article treat automatic summarization with approaches that provide the ability to summarize texts for different languages. In the case of this article, it is about the English, Romanian and Russian languages. The paper contains both the description of the problem and different approaches already used by other researchers. Next, the data with which the automatic summarization experiments were carried out were described. The metrics with which we can evaluate the quality of the summarization result were presented. Finally, some thoughts were formulated regarding the results obtained in the experiment.

**Keywords:** elearning systems, text summarization, evaluation metrics, datasets, data analysis.

**MSC 2020:** 68T50, 68T05.

## 1 Introduction

Humanity in the 21st century has made a huge leap in computer science. Chat GPT is the best demonstration of this thesis. It would seem that lots of information on the Internet hampers human work with its variation and validity. Chat GPT has succeeded with this problem and proposed remarkable results. This opportunity throws light on content generation [1].

On the other hand, because of the enormous amount of information daily generated the manual completion of elearning system with

appropriate content it is a difficult task for every teacher. With these special possibilities of automatic content generation, it becomes an interesting idea to supplement elearning platforms with a part of these huge amount of information referring to specific topics. In the conditions when there is so much information on the Internet, the need to reduce and understand it, in a limited time, becomes important. The processes of text understanding and production are directly related to the creation of summaries. That is why making a consistent summary is an important approach to understand and to select the appropriate idea to be included in educational materials on elearning platforms.

**Automatic text summarization** is a technique that takes a source text and extracts the most crucial information, condensing it and tailoring it to the demands of the user or job. The source text is first read, and its content is identified. The main points are then collected in a brief summary [2, pp. 2-4]

Searching approaches on automatic summarization, literature review brought us three solutions: *prompt engineering*, *abstractive-based summarization* and *extraction-based summarization*. The first two consider neural network technology, namely **transformers**. The last one relies on **standard NLP techniques** [3]. In this article, we consider the last solution: extraction-based summarization.

In contrast to abstractive techniques, which conceptualize and paraphrase a summary, extractive techniques accomplish summarization by selecting bits of texts and creating a summary [4].

The **purpose** of this article is to find the way of evaluating text summaries in the first place, to identify the best approach for summarization in the second place, and to investigate whether there are problems from a multilingual perspective in this procedure in the third place.

To achieve the goal of the paper, we will structure the paper as follows. Initially we will present the data we will work with, namely their type, quantity and scope. We will continue with the presentation of the methods and metrics needed to evaluate the experiment and after that the essence of the experiment and the data obtained will be presented. Finally, we will draw some conclusions based on what we obtained.

## 2 Overview of project articles

Content generation task can be viewed from different angles. Starting from the idea of adaptive content generation for eLearning platforms, the following can be regarded:

1. answers for student questions;
2. e-course content for teachers;
3. items from the test.

The last one refers to **adaptive assessment**. From this perspective, *the responsibility of item ordering is assigned to the software part*. There are two main categories of strategies for presenting test items: **two-step** and **multi-step** (Fig. 1) [5].

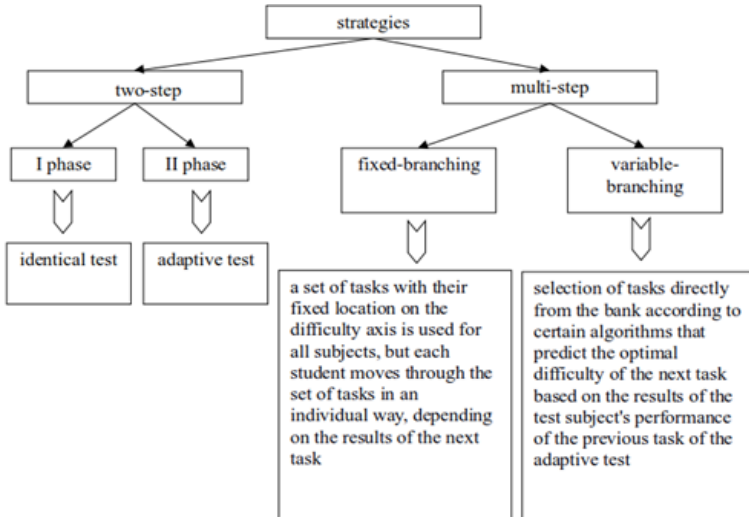


Figure 1. Adaptive testing strategies

As part of our project, it was decided to use the **Moodle** platform to implement our ideas and developments. Thus, we have developed two plugins: **TestWid** for adaptive assessment and **TestWidTheory** for content generation. The TestWid plugin is based on a **multi-step**

**fixed-branching strategy.** It uses a **bank of items** and **categories of items**. Let us group the items in these categories according to their complexity. Hereby, each time the student takes a test, new items are randomly selected (15 items in total). The plugin also allows you to **make retakes** for the same test with only one requirement: *there should be at least two items in each category*. So the student could obtain new items to deal with. Otherwise, the test could not be launched [5].

Another look at content generation regards **e-course development**. In contrast with adaptive assessment, where the notion of “generation” is treated as the order change of items, e-course development refers to the *fetching of information from the Internet in an advanced mode*. Our research papers [6],[7] suggest a focused web crawler based on a web-scraping approach for information extraction from the Internet and its further processing. Fig. 2 can provide a detailed view of our application for e-course development.

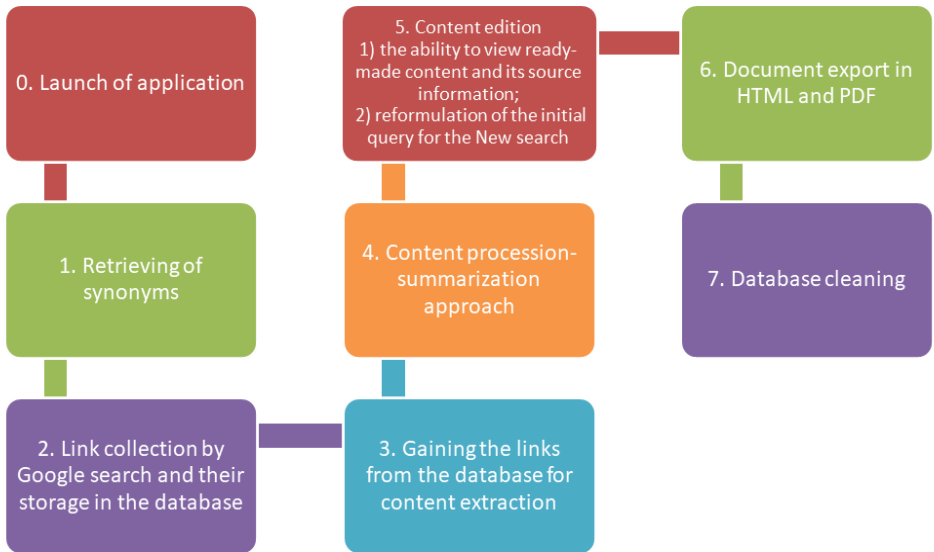


Figure 2. Scheme of the program model for the dynamic creation of e-courses

According to our approach, we have six steps. **In the first step,**



some web crawlers create networks of synonyms. **In the second step**, our application uses the original request and/or their selected synonyms for advanced search using Google search. **Next**, in step three, we gain links for the requests from Google and process them (crawl, select necessary fragments), storing all the information in the database. **Step 4** is responsible for the text processing and the extraction of the most valuable information. A literature review in this domain sheds light on the summary approaches. Summarization will help us interpret large amounts of the sources and present this resume as one final product, i.e., we can summarize one or many (depending on the summary approach) text sources and provide it to the user. According to **step 5**, the summarized content should be edited and further exported in HTML or PDF formats in **step 6**.

**From the student point of view**, content generation can be considered a question-answering solution. Insofar as it represents the essential way of understanding the learning material throughout the querying, it is a good practice to provide such assistance.

In this case, the same application for e-course generation can be the **prototype with some little changes**. Especially now, with the appearance of ChatGPT and other similar chat-bots, it becomes profitable to have one specific AI solution on the same platform that could effectively manage learning resources on the platform and outside it, on the one hand, and provide some extra possibilities, on the other.

Finally, this application is going to be integrated into the Moodle eLearning platform as a part of the **TestWidTheory** plugin and some new one for student assistance (chat-bot). Thus, this solution will be part of the Moodle standard tool set, which will always be at the teacher's hands.

To begin with the experiment, it should be mentioned that the Internet contains an enormous amount of information that requires careful processing, selecting only valuable passages. The recent study led us to the **summary approaches** to process the information taken from the internet and present it to the teacher. It is going to be implemented **in step 4** and will be discussed further.

### 3 Experiment description

In order to examine the quality of the extractive summarization, six texts from different sources with different structures and domains of topic were selected (see Table 1).

Table 1. Descriptions of the selected datasets

	Domain	Language	Chars	Words
1	History	English	8599	1309
2	Geography	English	7939	1326
3	Biology	Russian	10323	1356
4	Literature	Russian	53328	6879
5	Informatics	Romanian	20579	3063
6	Law	Romanian	11780	1652

As part of our experiment, we investigated various types of methods for automatic summarization. Some of them are built on plain speculations and others are built on more complicated algorithms. The following summary methods described in paper [8] were investigated:

1. **Luhn’s Heuristic Method** - propose that the **significance of each word** in a document **signifies how important it is**. According to this theory, *sentences that contain more of the stop-words* (words with the highest frequency) than others *do not have a greater impact* on the document’s meaning [13].
2. **Edmundson Heuristic Method** - recommends **the use of a subjectively weighted mixture of features**. He took into account the features that were previously well-known and utilised in Luhn’s method, but he also included a few new features such as *cue words* and *document structure* [14].
3. **Latent semantic analysis (LSA)** - is a reliable algebraic-statistical technique that can **find synonyms in the text and subjects that aren’t mentioned clearly in the text**. LSA works by *breaking down the data into small, manageable spaces* [2, p. 1002].

4. **SumBasic algorithm** - produces summaries of **length n**, where **n** is the user-specified number of sentences.
5. **Kullback-Lieber (KL) Sum algorithm** - its goal is to identify a set of sentences whose length is fewer than **L words** and whose unigram distribution closely resembles that of the source text [11, pp. 522-523].
6. **Graph-based summarization (Reduction)** - employs a graph to rank the necessary sentences or words in our **unsupervised strategy**. The primary goal of the graphical method is to extract the most significant sentences from a single source.
7. **LexRank algorithm** - is also a method related to graph based approach. It uses the cosine similarity of TF-IDF vectors;
8. **TextRank algorithm** - is also a method related to graph based approach. It uses measure based on the number of words two sentences have in common (normalized by the sentences' lengths).
9. **Term Frequency method** - enlightens us as to *which terms are most frequently used* and sheds light on the *significance of particular terms* in a given text or group of papers. The length of each document varies, thus *it is likely that a term will appear more frequently in larger documents* than in shorter ones. In order to normalize term frequency, it is frequently divided by the total number of terms in the document. Other methods of normalizing word frequencies include using the average and maximum term frequencies found in a document.
10. **Term Frequency-Inverse Document Frequency (TF-IDF)** - is a commonly used method in NLP to assess the importance of words in a document or corpus. IDF is a weight that **represents a word's usage volume**. The lower the score, the more frequently it is used throughout documents. A text vectorization procedure converts words in a text document into significance numbers. The TF-IDF vectorization/scoring method, as the name suggests, *multiplies the Term Frequency (TF)* and

*Inverse Document Frequency* (IDF) of a word to determine its score [15].

The first eight approaches were applied from the *Sumy* library for text summarization. The term frequency method was examined from *NLTK* and *Spacy* library and TF-IDF approach was examined by *NLTK* and *Scikit-Learn*. Summing up, we have investigated twelve methods for text summaries.

In order to estimate the quality of each method, we used four metrics discussed in [2]:

- ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) - score component provides a unique viewpoint on the effectiveness of the system-generated summary by taking various linguistic and grammatical elements into account [9, p. 74]. It defines *how much of the words in reference summaries* appeared in the candidate summaries.
- BLEU - is based on the basic idea of comparison machine translations/summarization with those regarded to be accurate by humans. Each segment (mainly sentences) is being compared with a set of qualitative reference texts. The obtained scores are then averaged over the whole corpus to reach an estimate of the translation's/summarization's overall quality [10, p. 394];
- METEOR - overpasses previous metrics, taking into account **grammar and semantics**. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision [10, p. 394];
- F-score - also relies on precision and recall, but data are different. Precision represents the **number of sentences** taking place in both summaries divided by the number of sentences in the candidate summary. The basic way how to compute the F-score is to count a harmonic average of precision and recall.

Beyond the upper metrics, we have used the metrics provided by the *Sumy* library, which evaluates its own algorithms with *ROUGE*, *F-score* and *Unit overlap* metrics.

The experiment consisted of a process that had two loops: an outside loop for changing texts and an inside loop to change summary methods (Fig. 3).

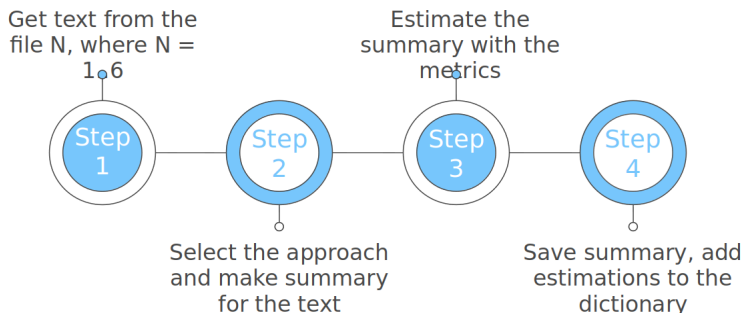


Figure 3. The course of the experiment.

At the end of the inner loop, all metrics were taken and analyzed. The experiment includes 72 iterations.

## 4 Data analysis

Working with the results, we have applied the **Max aggregated function** to get the highest results for each parameter. Our aim is understand which method is the best to be used and in what circumstances. That is why we analysed the methods from the following perspectives: overall effective summary approach; multilingual summary problems and approach versatility; comparison of approach realization and metrics confluence.

After applying the designed algorithm with methods to our datasets, the following results were obtained (see Table 2).

As it can be seen, from all the approaches, **the most effective are Luhn's heuristic method, TextRank and Term frequency method**. Having in mind **language sensitiveness**, we can use *Luhn's heuristic method for the English language* and *Term frequency for the Russian and Romanian languages*.

Table 2. Evaluation of the summary approaches. Part 1

<b>Rank</b>	<b>Language</b>	<b>ROUGE-1</b>	<b>ROUGE-2</b>	<b>ROUGE-L</b>	<b>BLEU</b>	<b>METEOR</b>
1	<b>En</b>	Luhn 0,8042	Luhn 0,9827	Luhn 0,8042	Luhn 0,5322	TF-IDF (NLTK) 0,59
2	En	TextRank 0,784	TextRank 0,74	TextRank 0,784	TextRank 0,499	LexRank 0,455
1	<b>Ru</b>	TF (Spacy) 0,86	TF (Spacy) 0,804	TF (Spacy) 0,86	TF (Spacy) 0,632	TF-IDF (Scikit-Learn) 0,524
2	Ru	TextRank 0,836	TextRank 0,793	TextRank 0,836	TextRank 0,584	Luhn 0,4761
1	<b>Ro</b>	TF (Spacy) 0,888	TF (Spacy) 0,822	TF (Spacy) 0,888	TF (Spacy) 0,686	TF-IDF (Scikit-Learn) 0,497
2	Ro	Luhn 0,8363	Luhn 0,7933	Luhn 0,8363	Luhn 0,5824	TF (Spacy) 0,427

**It should pay attention** to the METEOR results. ROUGE and BLEU results coincide, but METEOR's data differs. From all inputs, TF-IDF approach was frequently selected.

Another group of metrics is given below (see Table 3). Here F-score is based on the ROUGE and BLEU results and Unit overlapping. As we have three types of ROUGE metric in F-score formula we will get three types of F-score. Thus F-1 is for ROUGE-1, F-2 is for ROUGE-2 and, F-L is for ROUGE-L.

The last column of Table 3 (Unit overlapping) is calculated on the basis of **Summy** library that estimates *only its methods*. Thus not all summary approaches were taken into consideration.

Table 3. Evaluation of the summary approaches. Part 2

Rank	Language	F-1	F-2	F-L	Unit overlapping
1	<b>En</b>	Luhn 0,641	TF-IDF (NLTK) 0,626	Luhn 0,713	LSA/KL 0,33
2	En	TextRank 0,61	TF-IDF (Scikit-Learn) 0,564	TextRank 0,686	TextRank 0,30
1	<b>Ru</b>	TF (Spacy) 0,729	Reduction 0,591	TF (Spacy) 0,789	KL 0,38440
2	Ru	TextRank 0,688	Luhn 0,59	TextRank 0,755	Luhn 0,36705
1	<b>Ro</b>	TF (Spacy) 0,774	TF-IDF (Scikit-Learn) 0,59	TF (Spacy) 0,827	LexRank 0,36
2	Ro	Luhn 0,687	TF (Spacy) 0,562	Luhn 0,754	Luhn 0,33

Looking at the results, BLEU and METEOR provide an average value of around 50% of quality. This is comparable to 50% of the summarized volume of text. In contrast, ROUGE metric provided results about 80% of quality. This is normal because these metrics complement each other. You will have high BLEU if many terms from the candidate summary appear in the reference summary, and high ROUGE if many words from the candidate summary appear in the reference summary. The F-score, in this case, provides the common result as a summarization.

As a result, **the most effective approaches** from the second group of metrics are *Term frequency method*, *Luhn's heuristic method* and *Term Frequency-Inverse Document Frequency*. For English sources should be taken *Luhn's heuristic method* and *TextRank*, for Russian and Romanian languages suit *Term frequency*, *TextRank* and *Luhn's heuristic method*.

Unit overlapping metric emphasizes KL and Luhn's heuristic method in most cases. Unfortunately, this approach of evaluation was implemented in Sumy library for proper algorithms. And we cannot estimate other approaches.

Unfortunately, nothing can be said about the **readability** and **coherence** of the summaries. *The applied metrics cannot estimate these parameters*. Hypothetically, **as language is flexible** and has **different ways of expanding the context of speech**, there are problems **with its preservation during sentence extraction** during the extractive approach. For example, pronouns, link words, etc. Although we have noticed that each summary approach **cuts the original text at different places**, it is impossible to judge its efficacy without an expert review.

**The number of words**, it seems, **does not play an important role** in summary ranking. We had six texts of different lengths, but their summaries *were appreciated with the same high scores as others*. This is because we have indicated the summary length **in percentage**. That is why the ratio of the original text to its summary was always the same for metrics. **They pay attention only to the amount of corresponding words in both places**. That is not reasonable for extraction-based summarization. Though different scores show that



**some differences exist.**

Another thing that should be considered is **summary size**. The shorter the summary, the lower the quality. The results argue for good quality in both cases, with a small difference. More shorter summaries were not taken into account as we pursued the goal of designing **ecourse content in the educational area**. The courses should not be too small but contain relevant information for students.

The most productive summary approaches are Term frequency method, Luhn's heuristic method and TextRank. However, the language influence on method list seems to be vague. The **week point** in all these methods is **tokenization** and **stopwords** list. This topic necessitate more research and experiments for strong conclusions.

## 5 Conclusion

This paper is the extended and revised version of the conference paper [16] presented at WIIS 2023.

In this paper, we tried to find effective methods for text summaries from a multilingual perspective. All methods prefer the English language as the default. **Tokenization** and **stopwords** lists seem to **affect the Russian and Romanian languages**. Thereby, such "simple" approaches as TF or TF-IDF have high ranks compared to more advanced approaches.

It should be emphasized that this direction of research should be pursued. Now, the most effective methods are the term frequency method, Luhn's heuristic method, and TextRank.

The research has shown that **we cannot be firmly confident** in summary efficacy relying currently on evaluation metrics. We need some **expert opinion** to investigate such parameters as **readability** and **coherence** of the shortened texts. Only thereafter we can see what **pitfalls** also should be considered and conclude whether extraction-based summarization is good for **e-course content generation** and select the best approach. Also, we should regard some other solutions as **prompt engineering** and **abstractive-based summarization**.

**Acknowledgments.** This article was written within the framework of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

## References

- [1] T. Mali and R. Deshmukh, "Use of chat gpt in library services," *International Journal of Creative Research Thoughts*, vol. 11, no. 4, pp. f264–f266, Paper ID - 234893, 2023, [Online]. Available: <http://www.ijcrt.org/papers/IJCRT2304646.pdf>. DOI: 10.1729/Journal.33816.
- [2] Josef Steinberger and Karel Jezek, "Evaluation Measures for Text Summarization," *Computing and Informatics*, vol. 28, pp. 1001–1026, 2009. ISSN: 1335-9150.
- [3] Shashank Bhargav, Abhinav Choudhury, Shruti Kaushik, Ravindra Shukla, Varun Dutt, "A comparison study of abstractive and extractive methods for text summarization," *Advances in Intelligent Systems and Computing*, 2021, to be published.
- [4] S. Hima Bindu Sri and S. R. Dutta, "A Survey on Automatic Text Summarization Techniques," *Journal of Physics: Conference Series*, vol. 2040, no. 1, Article ID. 012044, 2021. DOI: 10.1088/1742-6596/2040/1/012044.
- [5] A. Parahonco and M. Petic, "How to create an adaptive e-learning system," in *Proceedings of the 17th International Conference "Linguistic Resources and Tools for Natural Language Processing"*, (Chisinau and Online 10-12 November 2022), 2022, pp. 153–160. ISSN 1843-911X.
- [6] A. Parahonco, M. Petic, and C. Negara, "The model of Web crawler for expansion the scope of initial search," in *Workshop on Intelligent Information Systems: Proceedings WIIS2020*, (Chisinau, 04-05 December 2020), 2020, pp. 139–150. ISBN 978-9975-68-415-6.

- [7] A. Parahonco and M. Petic, “Elearning content processing situations and their solutions,” in *Proceedings of Workshop on Intelligent Information Systems: WIIS2022*, (Chisinau, October 06-08, 2022), 2022, pp. 154–159. ISBN 978-9975-68-461-3.
- [8] Vishal Gupta and Gurpreet Lehal, “A Survey of Text Summarization Extractive Techniques,” *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, 2010. DOI: 10.4304/jetwi.2.3.258-268.
- [9] Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, (Barcelona, Spain), 2004, pp. 74–81.
- [10] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, “SummEval: Re-evaluating Summarization Evaluation,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021. DOI: 10.1162/tacl\_a\_00373.
- [11] S. Sripada and J. Jagarlamudi, “Summarization Approaches Based on Document Probability Distributions,” in *Pacific Asia Conference on Language, Information and Computation*, 2009, <https://api.semanticscholar.org/CorpusID:1150351>.
- [12] A. Li, T. Jiang, Q. Wang, and H. Yu, “The mixture of textrank and lexrak techniques of single document automatic summarization research in Tibetan,” in *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, IEEE, vol. 1, 2016, pp. 514–519.
- [13] S. Siddika and M. S. Hossen, “Automatic Text Summarization Using Term Frequency, Luhn’s Heuristic, and Cosine Similarity Approaches,” in *2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET)*, IEEE, 2022, December, pp. 1–6.
- [14] H. P. Edmundson, “New methods in automatic extracting,” *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

- [15] H. Christian, M. P. Agus, and D. Suhartono, “Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF),” *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, pp. 285–294, 2016.
- [16] Alexandr Parahonco and Mircea Petic, “Different aspects on extractive text summarization as a part of content generation for e-courses,” in *Proceedings of Workshop on Intelligent Information Systems: WIIS2023*, (Chisinau, October 19-21, 2023), 2023, pp. 169–180. ISBN 978-9975-68-492-7.

Alexandr Parahonco<sup>1</sup>, Mircea Petic<sup>2</sup>

Received September 30, 2023

Accepted December 12, 2023

<sup>1</sup> ORCID: <https://orcid.org/0009-0007-3486-5597>

Vladimir Andrunachievici Institute of Mathematics and Computer Science, Alecu Russo Balti State University

E-mail: [alexandr.parahonco@usarb.md](mailto:alexandr.parahonco@usarb.md)

<sup>2</sup> ORCID: <https://orcid.org/0000-0001-6044-7646>

Vladimir Andrunachievici Institute of Mathematics and Computer Science

E-mail: [mircea.petic@math.md](mailto:mircea.petic@math.md)

# Challenges associated with using AR technology in education

Inga Titchiev, Olesea Caftanatov, Dan Talambuta

## Abstract

As commonly understood, Augmented Reality (AR) technology, with its potential to blend digital content into the real world, has gained significant attention in various fields. However, its implementation is not without hurdles. Embracing AR comes with a set of challenges that cover technical, usability, ethical, and practical considerations. It may happen that the augmentation doesn't consistently yield the desired outcomes, work as initially intended, or deliver a seamless and satisfying user experience. In this article, we present the standards in the development and adoption of AR, also solutions to enhance augmented experiences based on our two years of developing educational applications.

**Keywords:** Augmented reality, artifacts, challenges.

**MSC 2020:** 97C70, 68T05.

## 1 Introduction

Augmented reality integration within the realm of education offers students the opportunity to engage in immersive experiences, thereby enhancing the learning process by making it more interactive, efficient, and meaningful. Augmented reality services within the educational and training domain empower users to interact with real-time applications and virtual elements that elucidate and illustrate concepts using multimedia, computer-based simulations, animations, quizzes, etc. This mode of augmented reality education effectively supplements conventional teaching and learning methodologies by fostering critical thinking, elevating student engagement, and enhancing their comprehension.

However, this augmentation doesn't consistently yield the desired outcomes, work as initially intended, or deliver a seamless and satisfying user experience. Various challenges and issues can arise [3] when working with augmented reality methods.

We consider that augmented learning is a personalized learning approach that adjusts to the needs of the learner. It offers real-time remediation to help learners better comprehend a subject, encouraging exploration and understanding [5]. This approach leverages technologies that incorporate multimedia and interaction, which researchers, teachers, pupils [6], and students [4] have enthusiastically embraced. Instead of emphasizing rote memorization, learners engage in an adaptive learning process that responds to their immediate context. Augmented content can dynamically adapt to the learner's surroundings and learning styles by presenting text, images, videos, or even audio, such as music or speech. Typically, this additional information is displayed through pop-up windows in computer-based environments.

Over the past two years, we have been actively exploring novel techniques and strategies for integrating augmented reality into educational settings. Drawing from our personal experiences of both setbacks and achievements, we have accumulated valuable insights that we are eager to share in this paper.

## 2 Standards in AR

First of all, an important point is about the role of standards in the development and adoption of emerging technologies like Augmented Reality (AR). Standards play a crucial role in ensuring interoperability, compatibility, and the overall development and acceptance of a technology. Here are some key insights and considerations regarding standards in AR:

1. **Early stage challenges:** AR is still in its relatively early stages of development, with both hardware and software evolving rapidly. This dynamic environment makes it challenging to establish concrete standards when the technology is still evolving and there isn't yet a clear consensus on what those standards

should be.

2. **Interoperability and compatibility:** Without standards, AR applications and devices may not work seamlessly with each other. This can hinder user experiences and limit the broader adoption of AR technology. Standardization can ensure that different AR devices and software can communicate effectively.
3. **Innovation vs. standardization:** There is often a tension between rapid innovation and the establishment of standards. Early on, the focus is on experimentation and pushing the boundaries of what's possible. Standardization can sometimes slow down this process. However, once the technology matures, standards become essential for widespread adoption.
4. **Industry collaboration:** The development of AR standards typically involves collaboration among industry stakeholders, including hardware manufacturers, software developers, and content creators. Various organizations, such as the Open AR Cloud, IEEE, and W3C, work on defining standards and best practices in the AR space.
5. **User privacy and security:** AR standards should also address privacy and security concerns, as AR involves real-world interaction and data collection. Setting standards for data privacy, security, and user consent is critical for ensuring the ethical use of AR.
6. **Consumer trust:** The lack of standards can create uncertainty and consumer hesitancy. When consumers don't know what to expect from an AR experience, they may be less likely to embrace the technology. Standards can help build trust and confidence in AR applications.
7. **Regulatory considerations:** As AR becomes more widespread, governments and regulatory bodies may step in to establish their own standards and guidelines for the technology, particularly in areas related to safety and public use. In our country this field is pretty new, so there are still no such regulations related to AR.

8. **Evolution of standards:** AR standards are likely to evolve over time as the technology matures. They may start with basic interoperability and compatibility standards and gradually expand to cover more aspects of AR, such as content creation, user interfaces, and hardware specifications.

While the absence of comprehensive AR standards is a current challenge, it's a natural part of the technology's early development. As AR technology continues to grow, it's expected that industry collaboration and evolving standards will contribute to its broader adoption and the creation of a more unified AR ecosystem. These standards will not only benefit developers and businesses but also provide a better, more consistent experience for AR users. In the next section, we'll present a few challenges and issues related to designing augmented markers and creating an augmented experience in itself.

### 3 Challenges and issues in creating augmented reality experiences

The integration of augmentation technology within the education field holds the promise of transforming the learning process for students, but it comes with its set of challenges. As this technology becomes increasingly accessible and sophisticated, it is crucial to contemplate the ramifications for educators and students. One of the most significant hurdles is the cost, as this technology can be prohibitively expensive, and educational institutions often operate within tight budgets. According to research [7], they organized interviews and the results shed light on several challenges associated with the implementation of handheld AR technology among school teachers. The primary obstacle identified was the lack of a universal Bring Your Own Device (BYOD) policy in primary schools. Some participants reported that their schools did not permit students to bring digital devices, and only a small percentage mentioned financial constraints among parents as a reason for this restriction. Concerns were raised by both parents and school administrators about the potential loss of students' devices, leading to a reluctance to allow personal devices in classrooms. Some



teachers expressed a preference for the school providing handheld devices in a controlled environment, such as a lab, to facilitate efficient use of AR for educational purposes.

Nevertheless, it is essential to weigh these initial costs against the potential long-term benefits of such investments.

Another challenge is the need for teachers to become proficient in using this technology to its fullest extent. Some of them may face troubles and require additional support to effectively utilize the technology. According to [8], numerous challenges have been identified in teachers' adoption of Augmented Reality. These challenges encompass the difficulty in delivering sensory-rich instruction, constrained class time, a limited understanding of AR technology, apprehension about potential technological setbacks, a shortage of necessary equipment, high associated costs, and inadequate training for handling technological tools. Additionally, teachers may harbor misconceptions about the utility of AR tools, leading to a lack of reliance on this technology, further impeding its integration into the teaching process. There exists a general lack of familiarity among teachers with AR technology, making them less inclined to engage in ad-hoc developments due to affordability concerns. The unawareness of effective strategies to enhance student learning motivation using AR tools is another challenge faced by teachers.

Additionally, teachers must remain vigilant regarding potential risks, including cybercrime and data security breaches that could arise from the use of human augmentation technology in the classroom.

Nevertheless, the most significant challenge lies in the development of augmented content itself. In our application's development, which relies on marker-based technology, we encountered the issues described in the following subsections.

### **3.1 Challenges related to markers and artifacts**

We consider markers to be the digital image of a trigger, and the physical one is called an artifact. According to [1], markers with ratings of 2-3 stars can pose problems during the recognition and tracking phases. Vuforia Engine relies on the grayscale version of markers to

identify features for recognition and tracking. If the image exhibits low overall contrast and a narrow, spiky histogram, it is unlikely to function effectively as a target image. In our initial set of markers, which were designed using black and white colors, markers received ratings of 2-3 stars after evaluation. As a response to these challenges, we addressed the issues by introducing color and additional features in the second batch of markers, consequently increasing their rating to 4-5 stars. Another set of challenges arose due to the similarity in marker designs, leading to confusion for AR cameras, as shown in Figure 1.

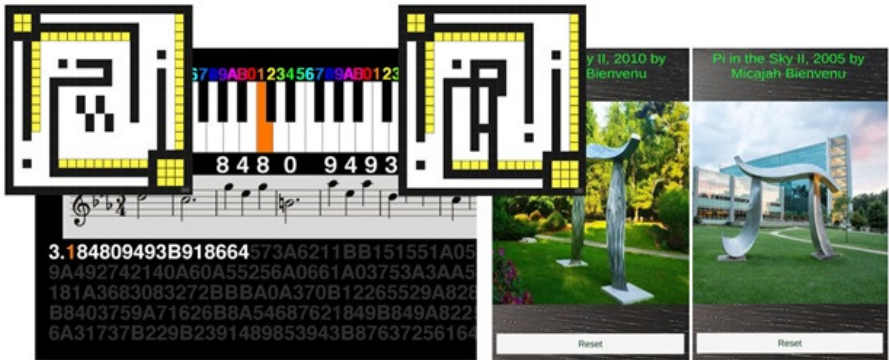


Figure 1. Example of similar markers that are confused

The left marker corresponds to the Pi Symphony augmentation experience, while the right marker randomly triggers the visualization of Pi in the Sky artworks by Micajah Bienvenu [2]. When using either one of these markers individually, AR cameras correctly identify and track the actions with 100% accuracy. However, when the markers were scanned sequentially, AR cameras triggered both marker scenarios simultaneously.

### 3.2 Challenges related to augmented learning content

A notable challenge pertains to the diversity of content to accommodate various learning styles and adhere to Bartle's Taxonomy. In order to provide students with personalized content that enables them to

learn more effectively, easily, and thoroughly through active engagement, it's imperative to identify areas for improvement. This entails affording students the opportunity to participate in refining the educational content and expressing both positive and negative feedback.

In this context, we conducted an experiment involving students from "Aleco Russo" University in Balti. The experiment involved providing participants with artifacts and mobile applications. Their objective was to test each artifact, which contained various types of augmented scenarios. Subsequently, we granted access to an online survey to gather information. The survey included questions related to the design of the artifacts, the augmented reality learning content, and the performance of the application itself. We collected data on the application's functionality, user satisfaction, and received recommendations for potential improvements. For example, some recommendations related to augmented scenarios are presented in Figure 2.

### **3.3 Challenges related to augmented scenarios**

Last year, we showcased some of our augmented artifacts at the International Exhibition of Creativity and Innovation, known as Excellent IDEA, which was organized by the Innovation and Technological Transfer Center of ASEM. During this exhibition, we received valuable recommendations for enhancing our scenarios. One particular scenario involved the creation of augmented artifacts for children with disabilities who were attending a children's camp. In this scenario, we aimed to provide an interactive experience showcasing 3D models of both wild and domestic animals along with their associated sounds, as shown in Figure 2 (Wolf v1.0). Subsequently, when we presented this augmented scenario to students from Aleco Russo University, they expressed interest, but they found it to be relatively straightforward. Their feedback prompted us to elevate the complexity of the scenario by introducing features that required user interaction with the 3D model. As a result, we incorporated four buttons that allowed the wolf to perform actions such as running, howling, lying down, and fighting, as depicted in Figure 2 (Wolf v2.0). Each action is followed by a sound.

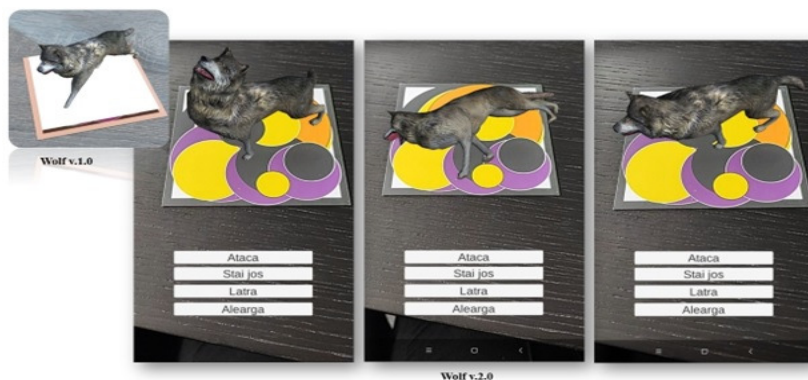


Figure 2. Example of improved scenario

## 4 Improvements of augmented reality experiences, solutions

Designing augmented markers and creating compelling augmented experiences come with their own set of challenges and considerations. Here are some summaries of the key challenges, issues, and solutions related to these aspects of augmented reality.

**Marker design and recognition:** Designing markers that are easily recognizable by AR systems while blending seamlessly into the physical environment can be challenging. Complex or cluttered markers may lead to recognition errors.

**Solution:** AR artifacts should be designed with clarity and intuitiveness in mind. Users should be able to understand the purpose and functionality of an artifact at first glance. The design should provide visual cues or information to guide users. If certain artifacts are not immediately intuitive, consider incorporating educational elements or tooltips to provide users with guidance on how to interact with them effectively. For example, adding quiz icons for artifacts that contain quizzes.

**Environmental variability:** AR systems can be sensitive to changes in lighting conditions. Creating markers and experiences that work well in various lighting environments, including low light and out-

door settings, can be a challenge. Our second version of artifacts was laminated with a glossy finish because it makes colors appear more saturated and imparts a professional and refined appearance. Moreover, glossy laminates excel in concealing fingerprints and smudges, simplifying the task of maintaining cleanliness. However, because of the highly reflective aspect, the AR experiences were difficult to recognize.

**Solution:** When it comes to laminating artifacts, a matte finish is the preferred choice. It effectively minimizes reflective light, and though it may slightly reduce color vibrancy, it maintains a subtle tactile quality that exudes professionalism. Although matte finish offers less protection compared to gloss and can show scratches and fingerprints, it doesn't pose concerns during recognition and tracking processes.

**User experience:** Ensuring that AR experiences are engaging and valuable to users is essential. It's crucial to strike a balance between the novelty of AR and the practicality of the experience.

**Solution:** One approach to achieve this balance is to actively engage teachers in the design of AR scenarios. Their input can help craft more effective and beneficial experiences that cater to educational needs.

**Content Creation:** Creating high-quality 3D models, animations, and interactive elements for AR experiences can be resource-intensive and require expertise in 3D design and development.

**Solution:** To enhance the quality and effectiveness of content, one effective solution is to enlist the services of a skilled designer, despite the potential cost involved. Alternatively, you can create your own 3D models and animations, which may be time-consuming but cost-effective. In our cases, we used free 3D models from Unity Assets.

**User interface (UI) design:** Designing user interfaces for AR experiences that are intuitive, non-intrusive, and accessible can be a complex task. The UI should complement the augmented content and not overwhelm users.

**Solution:** To address this challenge, we adopt a minimalist approach, striving for simplicity. Each artifact is dedicated to a single task or lesson, presenting information in manageable chunks. Additionally, for AR scenarios that involve user interaction during movement,

the design of artifacts is tailored to adapt to user actions, ensuring interaction without causing confusion.

**Tracking and calibration:** Accurate tracking and calibration are essential for a seamless AR experience. Maintaining the alignment of virtual objects with the physical environment, especially in dynamic situations, can be challenging.

**Solution:** Implement computer vision algorithms to improve object recognition and tracking accuracy. Feature detection, optical flow, and SLAM (Simultaneous Localization and Mapping) techniques can enhance real-time tracking. It is also possible to integrate machine learning algorithms to predict and compensate for potential inaccuracies in tracking. This can enhance the system's ability to adapt to various environments and user movements.

**Integration with real-world objects:** Incorporating real-world objects into AR experiences, like recognizing and interacting with specific physical objects, can be technically challenging due to variances in object appearance and shape.

**Solution:** Implement techniques such as feature detection, image matching, and deep learning to improve object recognition capabilities. Also can be created 3D models of real-world objects through scanning technologies (like photogrammetry or structured light scanning) to facilitate accurate digital overlays and interactions.

**Device accessibility and affordability:** One of the primary hurdles in integrating AR into education lies in the technology's technical prerequisites and constraints. For instance, AR demands devices like smartphones, tablets, or headsets that need to be both compatible and economically viable for teachers and pupils. The accompanying software, encompassing apps, platforms, or tools, must exhibit reliability, security, and user-friendliness for seamless integration. Moreover, the effectiveness of AR is contingent upon the quality and accessibility of internet connections, the durability of device battery life, and the adequacy of data storage capacity. Addressing these technical intricacies is essential to facilitate widespread and effective implementation of AR in educational settings.

**Solutions:**

- Governments can implement programs to provide subsidized or

free AR-compatible devices to schools, ensuring that both teachers and students have access to the necessary hardware.

- Schools can allocate budgets specifically for the purchase and maintenance of AR-enabled devices. Collaborations with technology providers may lead to discounted rates for educational institutions.
- Encourage BYOD policies [8] where students can use their personal smartphones or tablets for AR applications, minimizing the burden on schools to provide devices.

**Privacy and security:** AR apps often require access to a user's camera and potentially other sensors. Addressing privacy concerns and ensuring the secure handling of user data is a critical issue. Moreover, a growing trend involves the adoption of policies permitting students to use their personal devices at school. While this BYOD approach offers economic and operational advantages, it simultaneously introduces potential security vulnerabilities and negative consequences, contingent upon the ethical behavior of students and the absence of robust safeguards within school regulations [9].

**Solution:** Implement secure communication channels (e.g., SSL/TLS) to protect data integrity during transmission, using robust encryption protocols to prevent unauthorized access or interception, minimize data collection to only essential information required for AR functionality.

**Cross-platform compatibility:** Developing AR experiences that work seamlessly on multiple AR platforms (e.g., ARKit, ARCore, and various headsets) can be complex due to platform-specific requirements and capabilities.

**Solution:** Utilize standardized development frameworks like Unity 3D, Unreal Engine, PlugXR or WebXR that offer multiplatform support, allowing developers to create AR applications that work across different devices and operating systems. Conduct thorough testing across a range of devices, operating systems, and AR platforms to identify and address compatibility issues, ensuring a consistent user experience.

Addressing these challenges requires a combination of technical expertise, user-centric design, and a thorough understanding of the capabilities and limitations of AR technology. As AR continues to evolve, these challenges will also evolve, and new solutions will emerge to overcome them.

## 5 Conclusion

Furthermore, the introduction of human augmentation technology may alter the traditional dynamics of student learning and interaction. For instance, if students rely on augmented reality glasses to access educational content, it could reduce face-to-face interactions. While this technology has the potential to enhance learning outcomes, educators must be mindful of its potential impact on student engagement and socialization. With the introduction and adoption of smartphones and later introduction of Hololens and Oculus Rift, Augmented Reality technology that once seemed a thing of the somewhat distant future became feasible and started to evolve. It took some time to get a grasp on fundamentals but now the development of an augmented reality application is not much of a problem – just a matter of figuring out what’s and why’s.

Based on our two years of developing educational applications, in this article, the standards in the development and adoption of AR, also solutions to enhance augmented experiences were presented.

This paper is the extended and revised version of the conference paper [10] presented at WIIS 2023.

**Acknowledgments.** Intelligent Information systems for solving ill structured problems, knowledge and Big Data processing project Ref. Nr. 20.80009.5007.22, has supported part of the research for this paper.

## References

- [1] O. Caftanatov, I. Titchiev, V. Iamandi, D. Talambuta, and D. Caganovschi, “Developing augmented artifacts based on learning style approach,” in *Proceedings of WIIS2022, Workshop on Intel-*



- ligent Information Systems*, (October 06-08, 2022, Chisinau, Republic of Moldova), 2022, pp. 89–103. ISBN: 978-9975-68-461-3.
- [2] *Micajah Bienvenu's sculptures*, [Online]. Available: <https://www.micajahbienvenu.com/>.
- [3] N. M. Alzahrani, “Augmented Reality: A systematic review of its benefits and challenges in e-learning contexts,” *Applied Sciences*, vol. 10, no. 16, pp. 56–60, 2020.
- [4] J. Cabero-Almenara and R. Roig-Vil, “The motivation of technological scenarios in augmented reality (AR): Results of different experiments,” *Appl. Sci.*, vol. 9, no. 14, 2019, Article ID: 2907. DOI: <https://doi.org/10.3390/app9142907>.
- [5] H. Ardiny and E. Khanmirza, “The role of AR and VR technologies in education developments: Opportunities and challenges,” in *2018 6th RSI International Conference on Robotics and Mechatronics (ICRoM)*, (Tehran, Iran), 2018, pp. 482–487. DOI: <https://doi.org/10.1109/ICRoM.2018.8657615>.
- [6] J. M. Sáez-López, R. Cózar-Gutiérrez, J. A. González-Calero, and C. J. Gómez Carrasco, “Augmented reality in higher education: An evaluation program in initial teacher training,” *Educ. Sci.*, vol. 10, no. 2, Article no. 26, 2020 DOI: <https://doi.org/10.3390/educsci10020026>.
- [7] N. Alalwan, L. Cheng, H. Al-Samarraie, H. Al-Samarraie, R. Yousef, A. Alzahrani, and S. Sarsam, “Challenges and Prospects of Virtual Reality and Augmented Reality Utilization among Primary School Teachers: A Developing Country Perspective,” *Studies in Educational Evaluation*, vol. 66, Article ID: 100876, 2020. DOI: [10.1016/j.stueduc.2020.100876](https://doi.org/10.1016/j.stueduc.2020.100876).
- [8] Rajan Amar Bahadur Pal and Dr. Minesh Ade, “Applications and Challenges of Augmented Reality in Education Sector: A Report,” *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 10, no. 7, 2022, <https://doi.org/10.22214/ijraset.2022.45183>.

- [9] Madhavi Dhingra, “Legal Issues in Secure Implementation of Bring Your Own Device (BYOD),” *Procedia Computer Science*, vol. 78, pp. 179–184, 2016, <https://doi.org/10.1016/j.procs.2016.02.030>.
- [10] Inga Titchiev, Olesia Caftanatov, and Dan Talambuta, “Improving augmented reality experiences for application development,” in *Proceedings of Workshop on Intelligent Information Systems: WIIS2023*, (Chisinau, October 19-21, 2023), 2023, pp. 224–232. ISBN 978-9975-68-492-7.

Inga Titchiev<sup>1,4</sup>,  
Olesia Caftanatov<sup>2</sup>,  
Dan Talambuta<sup>3</sup>

Received September 30, 2023

Accepted November 27, 2023

<sup>1,2,3</sup>Vladimir Andrunachievici Institute of Mathematics  
and Computer Science, MSU  
5, Academiei street, Chisinau, Republic of Moldova, MD 2028

<sup>1</sup>ORCID: <https://orcid.org/0000-0002-0819-0414>

E-mail: [inga.titchiev@sti.usm.md](mailto:inga.titchiev@sti.usm.md)

<sup>2</sup>ORCID: <https://orcid.org/0000-0003-1482-9701>

E-mail: [olesea.caftanatov@math.md](mailto:olesea.caftanatov@math.md)

<sup>3</sup>ORCID: <https://orcid.org/0009-0008-7742-8597>

E-mail: [dantalambuta@gmail.com](mailto:dantalambuta@gmail.com)

<sup>4</sup> Ion Creanga State Pedagogical University of Chisinau

# On some aspects of medical data quality

Constantin Gaidric, Galina Magariu, Tatiana Verlan

## Abstract

This paper examines the specific problem of the quality of medical data of patients when they are hospitalized for treatment through the lens of the general problem of data quality. The need to apply international standards in data quality is discussed, taking into account the specifics of the medical field and the national standards and regulations. This necessity is considered with respect to the problem of medical institutions switching to IT systems with electronic medical records and electronic health records. The specifics of the procedures for filling in the files of hospitalized patients are highlighted.

**Keywords:** medical data, data quality, medical information systems, electronic medical record, electronic health record.

**MSC 2020:** 92C50, 68P27.

## 1 Introduction

The problem of data quality is becoming more and more important in all areas of human activity, even more so in healthcare. Even though virtually all of society understands the need to move to a digital economy, data is not treated as an essential and precious asset.

Data quality is not a simple scalar measure; it is defined by several dimensions, which reflect certain aspects with special meanings for different users, depending on the purposes for which the data is examined. Evaluation of the quality of data consists in establishing a value for each dimension by which it is appreciated how an aspect, a characteristic quality, is achieved, that allows understanding or decision-making according to the proposed objective.

In most medical institutions (ambulances, clinics, hospitals, laboratories) the data of patients' medical investigations are collected. Depending on the institution's management system, they are contained in the databases to which users from those institutions or the network of medical communities have secure access.

The quality of data in medical institutions has two aspects: the first one refers to the implications for the patient (without correct data we cannot expect the correctness of the diagnosis, and, therefore, the treatment process and, ultimately, the patient's health); the second one refers to the efficiency of the institution's management that ensures its functioning adequately to the tasks within the limits of the available sources. In this paper, the first aspect is examined. The impact of poor data quality can affect decision-making on treatment tactics and has a strategic influence on the treatment duration.

One of the most valuable assets in current business, as well as in planning for enterprises and institutions, is data. High-quality data is essential for each individual also. Data quality is naturally an evolving concept.

However, the issue of data quality is extremely important for correct decision-making and needs to be taken seriously. One example of this attitude is the UK, where this problem is considered at the government level. The Government Data Quality Framework was created, which provides "a more structured approach to understanding, documenting and improving the quality of its data" [12] (about data to which public bodies have access). "It presents a set of principles for effective data quality management, and provides practical advice to support their implementation." [12] In the development of this important framework as part of the National Data Strategy, the Data Management Association of the UK (DAMA UK) took part. All this shows the level of seriousness in approaching the problem of data quality in the UK.

The Data Management Association of the United Kingdom (DAMA UK) considers data quality dimensions to be "measurable features or characteristics of data" [12]. They are used to assess data quality and identify problems related to data quality, and therefore, the quality of important decisions. DAMA UK defines six basic dimensions of data quality (Completeness, Uniqueness, Consistency, Timeliness, Validity,

Accuracy), which can be and are completed according to user needs.

Understanding what good data means and how it can be measured and improved, if necessary, is quite a difficult problem for several reasons. There are a variety of definitions, and the number of dimensions taken into account differs from one domain to another depending on the context. Also, it depends on the vision of those who manipulate the data and for what purpose the data is applied [1].

Quality properties related to data usefulness are called “dimensions” in the literature on data quality.

A dimension is a measurable property of quality that represents some aspect of data (relevance, accuracy, consistency, etc.) and can be used to judge quality. Thus, some concrete data may be considered of high quality in certain respects according to one set of dimensions and less qualitative according to another set of dimensions. Probably, completeness, if this is not one of the most important dimensions, without which it is impossible to talk about the quality of some data, then this is the most frequently requested and encountered one. ISO 8000 [2] is the generally accepted standard for data quality in businesses and organizations.

In the last decades, in scientific publications and those of practical applications in management, increased attention is paid to the problem of the quality of data and information in databases. It is an undeniably important fact, but no less important is the quality of the data used in hospitals in the initial period, at the admission of the patient when a local database is formed from the data contained in the referral form from the family doctor and in the examination of the doctor in the admission department.

Data is fundamental to making correct, effective, evidence-based decisions. Perfect data quality is not always achievable, and therefore, the decision maker, in our case the physician, should understand what additional data would be needed to still ensure the achievement of the possible goal. So, a structured approach is needed to understand, document, and improve the quality of the data we intend to use.

According to many authors including [2], quality in use is generally considered the degree to which a product or a system can be used by users to achieve objectives with effectiveness, efficiency, lack of risk,

and satisfaction in specific contexts of use. The properties of quality in use are classified depending on the specifics of the field of activity through different sets of characteristics, among which the most common are: timeliness, precision, traceability, effectiveness, efficiency, and availability.

The standard ISO/IEC 25010:2011 “Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models” [2] defines:

- A model of quality in use and interaction of the product in a specific context, applicable to the human-computer system, including the mode of operation of the system.
- A product quality model composed of eight characteristics (which are subdivided into sub-characteristics) that relate to the static properties of software and dynamic properties of the information system.

The features in both models are relevant to both software products and any computer systems in the field. Terminology of characteristics (Accuracy, Completeness, Reliability, Relevance, Timeliness) and sub-characteristics is defined so that to be applied for measuring and evaluating quality (see Table 1). Thus, the set of quality characteristics can be selected for each context and compared if it corresponds to the applied standard.

We insistently promote the idea of using the ISO/IEC 25010:2011 standard even for traditional medical records (on paper), because their gradual transition to electronic support is inevitable. Information systems for electronic medical records keeping can ensure the high quality of medical data, their safety, accuracy, reliability, and easy exchange of information between patient and physician.

Nowadays, there is a world tendency of using patients’ medical data in different healthcare institutions in electronic format. According to data from The Commonwealth Fund [10], in developed countries, primary care physicians actively use this format in their practice: Australia – 97% of physicians; Canada – 86 %; England, Netherlands, and

Table 1. Characteristics of data quality in the ISO standard. Definitions adapted and modified according to [3]

<b>Characteristic of data quality in the ISO standard</b>	<b>Definition</b>
Correctness	the degree to which the data correctly represents the true value of the attribute
Completeness	data has values for all attributes expected in a specific usage context
Consistency	the degree to which the data is consistent and free of contradiction with other data
Credibility	the degree to which the data is considered true and credible by the user
Timeliness	the degree to which the data is age-appropriate in a specific context of use
Accessibility	data can be accessed in the context used
Conformity	the degree to which the data complies with the standards or regulations in force at the institution that maintains an information system
Confidentiality	the extent to which data is accessible and interpretable only by authorized users
Efficiency	the degree to which the data can be processed and provide the expected levels of performance
Precision	the degree to which the data is accurate
Traceability	the degree to which data access and changes are ensured
Understanding, comprehension	the degree to which data can be read and interpreted
Availability	the degree to which the data can be accessed
Portability	the degree to which the data allows installation, replacement or migration from one system to another maintaining the existing quality in a specific context of use
Recovery	the degree to which a certain level of operations and data quality is allowed to be maintained and preserved, even in the event of failure

Norway – 99 %; France and Germany – 88 %, Sweden – 98 %, New Zealand – 100 %, United States – 91 %.

Developing countries show a lower percentage of such use because of the problems in the level of health care and understanding the new technologies and their potential in resolving challenges in the domain [11].

We believe that in view of the transition over time of any system for patient records maintenance to the electronic version, it can be recommended to maintain and use these standards for patient records, a fact that would not require changes in the process of transition to electronic records.

Standard ISO/IEC 25012 [4] defines the data quality model that includes the same 15 characteristics and sub-characteristics from two points of view: inherent, which includes all the intrinsic characteristics, with the potential to explicitly and implicitly conform to the needs under the specified conditions; and dependent on the system, which refers to how data quality is obtained and maintained in an information system under specified conditions.

Decisions accompanying our daily lives should be based on high-quality data, so estimating data quality is a fundamental element in ensuring the relevance of decisions based on the data used. To increase confidence in data-driven decisions, it is necessary to measure and know the quality of the data used with appropriate tools [5].

A wide variety of commercial, open-source, and academic data quality tools have been developed based on scientific research. The range of functions offered by these tools varies widely, as the term “data quality” is context-dependent and not always used consistently.

Some tools exclusively offer data cleaning and enhancement functionalities that specifically address measurement capabilities, i.e., detection of quality issues and, most commonly, automated data modification (e.g., data cleaning). But this is not usually possible in information systems.

According to [8], tools that detect and report data quality problems are needed and a large number (667) are found, of which 50.82% are domain-specific, dedicated to certain types of data. Among the most common services they offer are data profiling, data quality measure-



ment, and continuous data quality monitoring, but these services are applicable only in specific areas.

## 2 Quality of data included in medical records

In healthcare, in recent years, many countries have transitioned or are in the process of transitioning from paper to digital records by hospitals, doctors' offices, clinics, and health care facilities. In the specialized literature, when talking about medical information systems, sources of data, and information about the patient, as a rule, the terms *electronic medical record* and *electronic health record* are used.

Electronic medical record (EMR) is a digital version of a patient's paper medical record that contains the patient's limited medical history completed mainly by the family doctor and the specialist doctors to whom the patient was referred for diagnosis and treatment. Patients' EMRs are typically owned and completed by primary institutions of specialized medical care, regardless of size. EMRs are not transmitted outside the institution unless the patient is admitted for inpatient treatment.

The electronic health record (EHR) contains patient information from all medical institutions that, over time, have been involved in patient care.

The EHR can include medical history, vital signs, progress notes, diagnoses, medications, immunization data, allergies, laboratory data, and imaging reports. It may also contain other relevant information such as insurance information and demographics. When talking about health care reform, it is necessary to emphasize the meaningful use of EHR. EHRs are designed as interoperable systems, which allow data from different systems to be accessed and used. Each medical institution can have access to the complete medical history of the patient, even if he was treated by other institutions. So the patient's medical information reaches every specialist, laboratory, or center, to which he calls. The EHR should be a comprehensive source of medical information on the general health status of patients, designed to be accessible by any authorized institution for diagnosis and medical care.

In Moldova, there are some popular EHR implementations and

many small EMR medical information systems installed in both private and public hospitals and diagnostic centers [6].

An example of an EHR is the DICOM Network. This is a distributed medical image preprocessing and archiving system.

The DICOM network was launched in Moldova in 2012 with the aim of providing access to collected imaging data for medical personnel with access rights and also access for patients when they need their personal data. Today, the system is implemented in many hospitals in Moldova, collects and processes more than 5 TB of data per month collected from hospitals equipped with different types of medical equipment [7].

Most of these systems contain both the data from the medical records of the patients registered at these medical institutions and also different types of medical image collections. The patient's personal data is the most sensitive and important information that should correspond to the main dimensions of data quality (completeness, correctness, timeliness, accessibility, conformity, etc.) in order to be used with confidence by each medical institution to which the patient has referred, for external consultations, when medical analyzes or images should be transferred to another medical institution, or when the patient wants to familiarize himself with the analyses and images of his latest investigations. A mandatory condition is securing access to any data and images from the EHR.

In the medical institutions of the Republic of Moldova, personal data is protected by the regulation NCPPD (National Center for the Protection of Personal Data of the Republic of Moldova), which is based on national legislation.

In most hospital institutions, internal standards for maintaining data from patient records are established. For example, in the Timofei Moșneaga Republican Clinical Hospital (Chisinau) in order to standardize data from different departments and to establish a standard method of completing and maintaining the inpatient medical record, the *Standard Operating Procedure* was approved (<https://scr.md/page/ro-proceduri-operaionale-269>). One can talk about pragmatic quality which refers to how the data enable the medical staff to achieve their goals and how easy to understand, how clear, how usable they are.

This procedure aims to ensure the quality of patient data and, first of all, the correctness and completeness of the medical care process at all stages in order to ensure the quality of the treatment given to the patient during his hospitalization and monitoring.

The most sensitive information present is medical data about patients. Denise Silber [7] finds that more than 86% of eHealth data errors are administrative or process errors: wrong recording of patient information, wrong test orders, and wrong drug prescriptions. The conclusions are based on a study conducted by 42 doctors over a 20-week period in 2000, which highlighted errors that occur outside the hospital setting.

Data quality is an important concern in any application scenario being both frequent and potentially costly. It is estimated that 2% of customer records become obsolete in a month due to data “degradation”. Medical centers are plagued by unresolved data quality issues. For medical centers, the most pressing need is to estimate data completeness.

Patients’ health can be adversely affected by a lack of documentation on treatments and medications. Other data quality considerations are not to be neglected. These domain-specific problems stem from deficiencies in data management processes or technical restrictions that may also occur in other domains. Data quality, in general, depends on context, that is, on notions of “good” or “poor” data that cannot be separated from the context in which they were produced and used (medical analysis data can be considered good as long as they fall within the limits for each type of analysis, for example, temperature 35-43°C, blood sugar – in the range of 70-120 mg per 100 ml).

### **3 Some procedures for filling in the records of inpatients**

The doctor from the inpatient department, after examining the patient and data from the accompanying documents, enters the *inpatient diagnosis* in the medical record of the inpatient.

Upon inpatient admission, the nurse at the registry completes the

inpatient's medical record, in which the *referral diagnosis* is entered according to the referral extract from the outpatient's medical record and the *inpatient diagnosis* (or *Admission Diagnosis* – diagnosis given to patient on admission to hospital).

The attending physician prescribes the exhaustive examination of the patient according to the admission diagnosis, and after the first 72 hours of admission specifies the diagnosis.

The medical record of the inpatient is completed by the staff from different subdivisions of the institution, who fix the information related to the diagnostic process, the dynamic evolution of the pathological process, and the applied treatment. The data in the record are intended to exhaustively reflect the assistance provided, the correctness and completeness of the data about the doctors' prescriptions, and the care provided by the nurses.

Given that the correct diagnosis and care of the patient are directly influenced by the quality of the data in the patient record, we will try to evaluate at what moments and on whom the quality of the inscriptions in the record depends.

Currently, not all medical institutions maintain patient records in electronic format and there are no databases of medical institutions connected in a functional national network.

The trends of managers and healthcare workers in the digital transformation of services and the use of artificial intelligence can only be achieved if they are based on good-quality data throughout the entire process: *preliminary diagnosis – hospitalization – evidence-based diagnosis – treatment – recommendations for the discharged patient and the family physician – storing data and information in the EHR* for subsequent decision-making and service delivery.

However, it is imperative to take into account the ISO/IEC 25010 standard that replaced ISO/IEC 9126-1:2001 and was developed to maintain the quality and evaluation process of the IT product. The standard includes a quality model in use composed of five characteristics (effectiveness, efficiency, satisfaction, freedom from risk, and context coverage) and a quality model defined by the ISO/IEC 25012 standard which includes 15 characteristics valid for any field. We will operate with only those specific to inpatient records for specifying,

measuring, and evaluating quality.

Although the scope of the product quality model is for IT systems, many of the characteristics are also relevant to wider systems and services.

The first step in completing the inpatient record is to enter the *referral diagnosis* and *admission diagnosis*.

That is why the data used to argue the referral diagnosis will be checked in terms of validity (if the defined parameters are according to the format of the given hospital and fall within the time interval in which the analyses are considered current), completeness (if they are sufficient to give credibility to the referral diagnosis), accuracy/ up-to-date (if they are not older than 3 days, so they can be used), compliance (the standards or regulations in force regarding the specific context of the case are respected), correctness (the data correctly represent the value of the attributes what confirms the diagnosis) and credibility (the degree to which the staff trusts the included data and reflects the doctor's needs in establishing the diagnosis).

For admission diagnosis, compliance is checked. The data included in the patient record are those that allow the attending physician to initiate the treatment process and prescribe his examination according to the protocol.

The analyses data and the attending physician's own observations lead to specifying the patient's diagnosis based on which the treatment continues.

The attending physician must be sure that the data he accesses and based on which he will propose the tactics of the patient's treatment are sufficient and satisfactory to obtain a result in a reasonable time.

Since the attending physician acts according to the protocol, and all investigations including analyses are carried out following the internal regulations of the hospital, the problem of trust in the quality of the data used to specify the diagnosis and the treatment of the patient becomes less acute. When the patient is discharged, it should be determined whether the data used by the doctor should be archived. In such a case, they must be: 1) accessible to the medical institutions to which the patient later turns and 2) secured. It would also be useful to store the data together with the information about its quality. How-

ever, in order to ensure traceability and interoperability of the system and compliance with all the compartments of the ISO 8000 standard, it is necessary to monitor the elements that characterize such dimensions as accuracy, promptness, conformity, and completeness.

## 4 Conclusions

This paper is the extended and revised version of the conference paper [13] presented at WIIS 2023. In this article, the problem of the quality of medical data was examined with the aim of facilitating the transition process to integrated health data systems that would provide the possibility for the attending physician to dispose of all the necessary medical investigations of the patient regardless of the location and the medical unit (laboratory, clinic) at which they were performed.

**Acknowledgments.** The project Ref. Nr. 20.80009.5007.22 “Intelligent Information systems for solving ill structured problems, knowledge and Big Data processing” has supported part of the research for this paper.

## References

- [1] C. Batini, M. Palmonari, and G. Viscusi, “Opening the closed world: A survey of information quality research in the wild,” in *The Philosophy of Information Quality*, Springer International Publishing, 2014, pp. 43–73.
- [2] ISO/IEC 25010:11. *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models*, 2011. [Online]. Available: <https://www.iso.org/obp/ui/#iso:std:iso-iec:25010:ed-1:v1:en>.
- [3] Tuomo Uotila and Helina” Melkas, “Complex Knowledge Conversion Processes and Information Quality in Regional Innovation Networks,” *Knowledge and Process Management*, vol. 15, no. 4, pp. 224–234, 2008, [www.interscience.wiley.com](http://www.interscience.wiley.com). DOI: 10.1002/kpm.317.

- [4] ISO/IEC 25012 [ISO/IEC, “Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model,” ISO/IEC, Tech.Rep. ISO/IEC 25012, 2008.
- [5] Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz, “Requirements for Data Quality Metrics,” *Journal of Data and Information Quality*, vol. 9, no. 2, pp. 12:1–12:32, January 2018.
- [6] Alexandr Golubev, Petru Bogatencov, Grigore Secrieru, Ecaterina Matenco, “Incident Handling and Personal Data Protection in Medical Images systems,” in *Intelligent information systems for solving weakly-structured problems, processing knowledge and big data*, Chisinau, Editors: S. Cojocar, C. Gaindric, I. Tițchiev, and T. Verlan. Chișinău: Vladimir Andrunachievici Institute of Mathematics and Computer Science, 2022, pp. 223–228. ISBN 978-9975-68-415-6. 004+519.7(082).
- [7] Peter Bogatencov, Nicolai Iliuha, Grigore Secrieru, and Alexandr Golubev, “DICOM Network for Medical Imagistic Investigations Storage, Access and Processing,” *Networking in Education and Research Proceedings of the 11th RoEduNet IEEE International Conference*, (Sinaia, Romania), January 17-19, 2013, pp. 38–42.
- [8] Lisa Ehrlinger and Wolfram Wöß, “A Survey of Data Quality Measurement and Monitoring Tools,” in *Front. Big Data*, vol. 5, Sec. *Data Mining and Management*, 31 March 2022, <https://doi.org/10.3389/fdata.2022.850611>.
- [9] Guide to Health and the Best Doctors. [Online]. Available: <http://www.hbroussais.fr/HEGP>. (in French)
- [10] The Commonwealth Fund. “Primary Care Physicians’ Use of Electronic Medical Records, Selected Health & System Statistics,” June 05, 2020. [Online]. Available: <https://www.commonwealthfund.org/international-health-policy-center/system-stats/primary-care-physicians-emrs>.

- [11] Mosharop Hossian, “Electronic Medical Record and Electronic Health Record have potential to improve quality of care,” *ACADEMIA Letters*, July 2021. [Online]. Available: <https://doi.org/10.20935/AL1565>.
- [12] Government Data Quality Hub, “Guidance. The Government Data Quality Framework,” gov.uk, 3 December 2020. [Online]. Available: <https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework>.
- [13] Constantin Gaidric, Galina Magariu, and Tatiana Verlan, “On some aspects of medical data quality,” in *Proceedings of Workshop on Intelligent Information Systems: WIIS2023*, (Chisinau, October 19-21, 2023), 2023, pp. 128–140. ISBN 978-9975-68-492-7.

Constantin Gaidric<sup>1</sup>, Galina Magariu<sup>2</sup>,  
Tatiana Verlan<sup>3</sup>

Received August 24, 2023  
Accepted November 29, 2023

<sup>1,2,3</sup>Vladimir Andrunachievici Institute of Mathematics  
and Computer Science, SUM

5, Academiei street, Chisinau, Republic of Moldova, MD 2028

<sup>1</sup> ORCID: <https://orcid.org/0009-0003-2893-9626>

E-mail: [constantin.gaidric@math.md](mailto:constantin.gaidric@math.md)

<sup>2</sup> ORCID: <https://orcid.org/0009-0001-4860-8715>

E-mail: [galina.magariu@math.md](mailto:galina.magariu@math.md)

<sup>3</sup> ORCID: <https://orcid.org/0009-0006-4519-1105>

E-mail: [tatiana.verlan@math.md](mailto:tatiana.verlan@math.md)



# On some classes of formulas in $S5$ which are pre-complete relative to existential expressibility

Andrei Rusu, Elena Rusu

## Abstract

Existential expressibility for all  $k$ -valued functions was proposed by A. V. Kuznetsov and later was investigated in more details by S. S. Marchenkov. In the present paper, we consider existential expressibility in the case of formulas defined by a logical calculus and find out some conditions for a system of formulas to be closed relative to existential expressibility. As a consequence, it has been established some pre-complete as to existential expressibility classes of formulas in some finite extensions of the paraconsistent modal logic  $S5$ .

**Keywords:** Paraconsistent logic, existential expressibility, logical calculi.

**MSC 2020:** 68R99, 68Q25, 06E25, 03B53.

**ACM CCS 2020:** 10003752.10003790.10003793.

## 1 Introduction

It is a well known class of problems in logic, algebra, discrete mathematics, and cybernetics dealing with the possibility of obtaining some functions (operations, formulas) from other ones by means of a fixed set of tools. The notion of expressibility of Boolean functions through other ones by means of superpositions goes back to the works of E. Post [1], [2]. He described all closed (with respect to superpositions) classes of 2-valued Boolean functions. The problem of completeness (with respect to expressibility), which requires to determine the necessary and sufficient conditions for all formulas of the logic under investigation to be expressible via the given system of formulas, is also

investigated. In 1956 ([3, p. 54], [4]), A. V. Kuznetsov established the theorem of completeness according to which we can build a finite set of closed with respect to expressibility classes of functions in the  $k$ -valued logics such that any system of functions of this logic is complete if and only if it is not included in any of these classes. In 1965 [5], Rosenberg I. established the criterion of completeness in the  $k$ -valued logics formulated in terms of a finite set of pre-complete classes of functions, i.e., in terms of maximal, incomplete, and closed classes of functions.

In the present paper, we investigate necessary conditions of completeness with respect to existential expressibility of the systems of formulas in some extensions of the modal logic  $S5$ .

The standard language of  $S5$  is based on propositional variables and logical connectives:  $\&$ ,  $\vee$ ,  $\rightarrow$ ,  $\neg$ ,  $\Box$ , and  $\Diamond$ . We consider the para-consistent negation  $\sim$  of  $S5$  [6] as follows:

$$\sim a =_{Def} \Diamond \neg a.$$

The logic  $S5$  can be considered, according to [6], as a para-consistent logic since it contains a para-consistent negation. The logic  $S5$  is characterized by the axioms and rules of inference of the classical propositional logic, the following axioms ( $A$  and  $B$  are any valid formulas):

$$\begin{aligned} \Box(A \rightarrow B) &\rightarrow (\Box A \rightarrow \Box B), \\ \Box A &\rightarrow A, \\ \Diamond A &\rightarrow \Box \Diamond A, \end{aligned}$$

and the necessity rule of inference: from  $A$  infer  $\Box A$ .

Consider the set  $E_k$  of finite binary strings  $(\alpha_1, \dots, \alpha_k)$ , where  $\alpha_i \in \{0, 1\}$ ,  $i = 1, \dots, k$ . Define Boolean operations  $\&$ ,  $\vee$ ,  $\rightarrow$ ,  $\neg$  over elements of  $E_k$  component-wise, and consider  $\Box((1, \dots, 1)) = (1, \dots, 1)$ , and put  $\Box((\alpha_1, \dots, \alpha_k)) = (0, \dots, 0)$  otherwise. Also, as usual,  $\Diamond x = \neg \Box \neg x$ . It is known [7] that  $(E_k; \&, \vee, \rightarrow, \neg, \Box, \Diamond)$  represents an algebraic model for  $S5$ .

Kuznetsov A. V. proposed in [8] some generalizations of the notion of expressibility of formulas in a superintuitionistic logic, namely the parametric expressibility, and the existential expressibility.

The formula  $F$  is said to be expressible in the logic  $L$  via a system of formulas  $\Sigma$  if  $F$  can be obtained from propositional variables, constants, and formulas of  $\Sigma$  applying a finite number of times: a) the rule of substitution of equivalent formulas in the logic  $L$ , and b) the rule of weak substitution, which permits, being given formulas  $A$  and  $B$ , to substitute one of them in another instead of a given corresponding propositional variable [8]–[10].

A. V. Kuznetsov [9] extended the notion of (explicit) expressibility from Boolean functions to formulas of the superintuitionistic propositional logics. He proposed the use of two rules (weak substitution and replacement by equivalent formula in the given logic) instead of the rule of superposition, and the problem of completeness with respect to (explicit) expressibility was solved for intuitionistic propositional logic and its extensions by M. F. Ratsa [10], [11].

In his work [8], A. V. Kuznetsov, among other things, extended the notion of (explicit) expressibility by modifying the tools previously used so as to obtain new formulas in superintuitionistic propositional logics and in the general  $k$ -valued logic  $P_k$ . Thus, he proposed the notions of implicit expressibility, parametric expressibility, and existential expressibility. The last one is similar to the notion of existential definability of predicates in arithmetics, examined by J. Robinson in [12].

Related to the problems of expressibility is the following one, which requires finding a tool (property)  $X$  that will permit us to separate the object  $A$  from the given system of objects  $\Sigma$  in the sense that if  $A$  is not expressible via the objects of  $\Sigma$ , then the objects of  $\Sigma$  possess property  $X$ , and the object  $A$  does not possess it. In this case, we speak about *separability of  $A$  from  $\Sigma$  by means of  $X$* , or we say that  *$A$  is detachable from  $\Sigma$  by means of  $X$* . In [8], A. V. Kuznetsov stated conditions of separability of a formula of the general  $k$ -valued logic from a given set of formulas with respect to explicit, parametric, and existential expressibility.

In the present paper, we specify the notion of existential expressibility to any algebra with a finite set of basic operations and we determine sufficient conditions for a system of term functions to be closed with respect to existential expressibility in the given algebra. As a consequence, some pre-complete relative to existential expressibility classes

of formulas in some tabular extensions of the logic  $S5$  are identified.

## 2 Basic notions

Consider the set of variables  $Var$ , whose elements will usually be denoted by small italic letters  $a, b, d, p, q, \dots$ , possibly with indices. Let  $\mathfrak{A} = (E; F_1, \dots, F_n)$  be an algebra with support  $E$  and basic operations  $F_1, \dots, F_n$ . The elements of the support of  $\mathfrak{A}$  are denoted by small Greek letters  $\alpha, \beta, \gamma, \delta, \dots$ . Terms of  $\mathfrak{A}$  are defined as usual [13, p.62, Def. 10.1] and are denoted by capital letters. In order to stress that the variables  $p_1, \dots, p_n$  occur in the term  $A$ , we will write  $A(p_1, \dots, p_n)$ . We will usually write the fact that some variable  $p$  is substituted in term  $A(p, p_1, \dots, p_n)$  by term  $B$  in the form  $A[p/B]$  or  $A[B]$  for short. The same notation  $A[p/\gamma]$ , or  $A[\gamma]$ , or  $A(\gamma)$  is used to denote the fact that the variable  $p$  is evaluated on  $\mathfrak{A}$  by the element  $\gamma$  of  $E$ .

The set of variables occurring in the term  $F$  is denoted by  $Var(F)$ . The set of terms of  $\mathfrak{A}$  is denoted by  $Term(\mathfrak{A})$  or shortly by  $Term$  (if there is no danger for confusion). The equality  $\mathfrak{A} \models A \approx B$  of 2 terms  $A$  and  $B$  on  $\mathfrak{A}$  is defined as usual [13], i.e., for any evaluation of variables with elements from  $E$ , the values of the terms  $A$  and  $B$  coincide.

**Definition 1.** (compare with [8]) *The term  $F \in Term(\mathfrak{A})$  is said to be expressible via the system of terms  $\Sigma$  on  $\mathfrak{A}$  if it is equivalent to a term  $G$  of the algebra  $(E; \Sigma)$  on  $\mathfrak{A}$ .*

Consider first-order formulas over  $Terms$  on  $\mathfrak{A}$  as usual, based on first-order connectives  $\approx, \vee, \wedge, \rightarrow$ , and  $\neg$  (equality, conjunction, disjunction, implication, and negation) and quantifiers  $\forall$  and  $\exists$ , respectively. Let  $\Psi$  be a first-order formula. The usual fact that  $\Psi$  is valid on  $\mathfrak{A}$  will be denoted by  $\mathfrak{A} \models \Psi$  or simply by  $\models \Psi$ .

**Definition 2.** *A term  $A$  is said to be existentially expressible via the system of terms  $\Sigma$  on  $\mathfrak{A}$  (see [8, p. 30] and take into consideration [8, p. 25]), if there exist: a) integer positive numbers  $l, m$ , and  $k$ ; b)  $\pi, \pi_1, \dots, \pi_l \in Var \setminus Var(A)$ ; c)  $B_{ij}, C_{ij}, D_t \in Term$  ( $i=1, \dots, m$ ;  $j=1, \dots, k$ ;  $t=1, \dots, l$ ) such that i)  $B_{ij}, C_{ij}$  are expressible via  $\Sigma$  on  $\mathfrak{A}$ ; ii)*

$\pi, \pi_1, \dots, \pi_l \notin \text{Var}(D_i)$  ( $i = 1, \dots, l$ ); and iii)

$$\models (A \approx \pi) \rightarrow (\bigvee_{j=1}^k \bigwedge_{i=1}^m (B_{ij} \approx C_{ij}))[\pi_1/D_1] \dots [\pi_l/D_l], \quad (1)$$

$$\models (\bigvee_{j=1}^k \bigwedge_{i=1}^m (B_{ij} \approx C_{ij})) \rightarrow (A \approx \pi). \quad (2)$$

**Example 1.** (compare with [8, p. 30]) Let us consider the Boolean algebra  $\langle \{0, 1\}; \&, \vee, \neg, 0, 1 \rangle$ , where  $\&, \vee, \neg$  are defined as usual. Then boolean functions  $p \& q$  and  $\neg p$  are existentially expressible via the constants 0 and 1.

We have

$$\begin{aligned} \models ((p \& q) \approx r) \approx & (((p \approx 0) \wedge (q \approx 0) \wedge (r \approx 0)) \\ & \vee ((p \approx 0) \wedge (q \approx 1) \wedge (r \approx 0)) \\ & \vee ((p \approx 1) \wedge (q \approx 0) \wedge (r \approx 0)) \\ & \vee ((p \approx 1) \wedge (q \approx 1) \wedge (r \approx 1))). \end{aligned} \quad (3)$$

According to [8, p. 30], we also have:

$$\begin{aligned} \models ((\neg p) \approx q) \approx & (((p \approx 0) \wedge (q \approx 1)) \\ & \vee ((p \approx 1) \wedge (q \approx 0))). \end{aligned} \quad (4)$$

The closure of the system  $\Sigma$  of terms relative to (existential) expressibility is defined as usual. In the present paper, only terms on  $\mathfrak{A}$  are considered.

**Definition 3.** A term  $A(p_1, \dots, p_n)$  is said to **conserve on  $\mathfrak{A}$  the relation  $R$**  (compare with [9]) if, for any elements  $\alpha_{ij} \in \mathfrak{A}$  ( $i = 1, \dots, n$ ;  $j = 1, \dots, s$ ), the facts  $\models R(\alpha_{i1}, \dots, \alpha_{is})$  imply  $\models R(F[\alpha_{11}, \dots, \alpha_{1n}], \dots, F[\alpha_{s1}, \dots, \alpha_{sn}])$ . Also, the system of terms  $\Sigma$  is said to **conserve the relation  $R$  on  $\mathfrak{A}$**  if any term of  $\Sigma$  conserves  $R$  on  $\mathfrak{A}$ .

### 3 Preliminary results

The next theorem provides sufficient conditions for a system of terms  $\Sigma$  to be closed relative to existential expressibility on  $\mathfrak{A}$ .

**Theorem 1.** *Suppose a)  $\mathfrak{A}$  is an algebra with an arbitrary finite set of operations; b)  $\mathfrak{A}_i$  are subalgebras of  $\mathfrak{A}$ ,  $i = 1, \dots, s$ ; c)  $\Phi$  be any mapping  $\Phi : \mathfrak{A}_i \rightarrow \mathfrak{A}$ ; d)  $K$  is a set of terms of  $\mathfrak{A}$  that conserve on  $\mathfrak{A}$  the relation  $R(y, x)$  of the type*

$$y = \Phi(x). \quad (5)$$

*Then  $K$  is closed with respect to existential expressibility.*

**Proof.** Let us suppose on the contrary that there exists a term  $A \notin K$  and  $A$  is existentially expressible via terms of  $K$ . Let  $a_1, \dots, a_n \in \text{Var}(A)$ .

Then, according to Definition 2, there exist a) terms  $B_{11}, C_{11}, \dots, B_{mk}, C_{mk}$ , and  $D_1, \dots, D_l$ ; b) variables  $a, d_1, \dots, d_l$  such that: i)  $B_{11}, C_{11}, \dots, B_{mk}, C_{mk}$  are expressible via  $K$  on  $\mathfrak{A}$ ; ii)  $a, d_1, \dots, d_l \notin \text{Var}(D_i)$  ( $i = 1, \dots, l$ ); and iii)

$$\models (A \approx a) \rightarrow (\bigvee_{j=1}^k \bigwedge_{i=1}^m (B_{ij} \approx C_{ij}))[d_1/D_1] \dots [d_l/D_l], \quad (6)$$

$$\models (\bigvee_{j=1}^k \bigwedge_{i=1}^m (B_{ij} \approx C_{ij})) \rightarrow (A \approx a). \quad (7)$$

We can consider in the following that  $a, a_1, \dots, a_n, d_1, \dots, d_l \in \bigcup_{i=1}^m \bigcup_{j=1}^k \{\text{Var}(B_{ij}) \cup \text{Var}(C_{ij})\}$ . So, since  $B_{ij}, C_{ij} \in K$  ( $i = 1, \dots, m$ ,  $j = 1, \dots, k$ ), i.e., they conserve relation (5), we have

$$\models \Phi(B_{ij}[\alpha_{11}, \dots, \alpha_{n1}, \alpha_1, \delta_{11}, \dots, \delta_{l1}]) = \left. \begin{array}{l} B_{ij}[\Phi(\alpha_{11}), \dots, \Phi(\alpha_{n1}), \Phi(\alpha_1), \Phi(\delta_{11}), \dots, \Phi(\delta_{l1})], \end{array} \right\} \quad (8)$$

$$\models \Phi(C_{ij}[\alpha_{11}, \dots, \alpha_{n1}, \alpha_1, \delta_{11}, \dots, \delta_{l1}]) = \left. \begin{array}{l} C_{ij}[\Phi(\alpha_{11}), \dots, \Phi(\alpha_{n1}), \Phi(\alpha_1), \Phi(\delta_{11}), \dots, \Phi(\delta_{l1})], \end{array} \right\} \quad (9)$$

where  $\alpha_{u1}, \alpha_1, \delta_{w1} \in \mathfrak{A}_i$ ,  $u = 1, \dots, n$ ;  $w = 1, \dots, l$ .

It follows from  $A(a_1, \dots, a_n) \notin K$  that  $A$  does not conserve relation (5). This means that there exist elements  $\beta_{u1}$ ,  $u = 1, \dots, n$  such that

$$\Phi(A[\beta_{11}, \dots, \beta_{n1}],) \neq A[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1})]. \quad (10)$$

Let us denote

$$A[\beta_{11}, \dots, \beta_{n1}] = \beta_1 \quad (11)$$

So, we have:

$$\models A[\beta_{11}, \dots, \beta_{n1}] \approx \beta_1. \quad (12)$$

Substituting (11) in (10), we get:

$$\Phi(\beta_1) \neq A[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1})]. \quad (13)$$

Let us expand the relation (6):

$$\begin{aligned} &\models (A(a_1, \dots, a_n) \approx a) \rightarrow \\ &(\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}(a_1, \dots, a_n, a, d_1, \dots, d_l) \approx \\ &C_{ij}(a_1, \dots, a_n, a, d_1, \dots, d_l))) [d_1/D_1] \dots [d_l/D_l]. \end{aligned} \quad (14)$$

The last relation takes place for any elements of  $\mathfrak{A}$ . In particular, we have

$$\left. \begin{aligned} &\models (A[\beta_{11}, \dots, \beta_{n1}] \approx \beta_1) \rightarrow \\ &(\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, d_1, \dots, d_l] \approx \\ &C_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, d_1, \dots, d_l])) \\ &[d_1/D_1[\beta_{11}, \dots, \beta_{n1}]] \dots [d_l/D_l[\beta_{11}, \dots, \beta_{n1}]]. \end{aligned} \right\} \quad (15)$$

Since relations (12) are true, we have from (15) the following:

$$\left. \begin{aligned} &\models (\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, d_1, \dots, d_l] \approx \\ &C_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, d_1, \dots, d_l])) \\ &[d_1/D_1[\beta_{11}, \dots, \beta_{n1}]] \dots [d_l/D_l[\beta_{11}, \dots, \beta_{n1}]]. \end{aligned} \right\} \quad (16)$$

Let us denote the elements  $D_w[\beta_{11}, \dots, \beta_{n1}]$  by  $\tau_{w1}$  for any  $w = 1, \dots, l$ . Then from (16) we have:

$$\left. \begin{aligned} &\models (\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{l1}] \approx \\ &C_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{l1}])) \end{aligned} \right\} \quad (17)$$

Let us look now at relation (7). For any elements  $\gamma, \gamma_1, \dots, \gamma_n \in \mathfrak{A}$ , we get:

$$\left. \begin{aligned} &\models (\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\gamma_1, \dots, \gamma_n, \gamma, d_1, \dots, d_l] \approx \\ &C_{ij}[\gamma_1, \dots, \gamma_n, \gamma, d_1, \dots, d_l])) \rightarrow (A[\gamma_1, \dots, \gamma_n] \approx \gamma). \end{aligned} \right\} \quad (18)$$

So, (18) is also true for particular elements  $\gamma, \gamma_1, \dots, \gamma_n \in \mathfrak{A}$ , where  $\gamma = \Phi(\beta_1)$ ,  $\gamma_1 = \Phi(\beta_{11}), \dots, \gamma_n = \Phi(\beta_{n1})$ , and we get:

$$\left. \begin{aligned} \models (\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), d_1, \dots, d_l] \approx \\ C_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), d_1, \dots, d_l])) \rightarrow \\ (A[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1})] \approx \Phi(\beta_1)). \end{aligned} \right\} \quad (19)$$

According to (13), we have:

$$\not\models A[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1})] \approx \Phi(\beta_1). \quad (20)$$

Then it follows that relation (19) holds if the next one is true:

$$\left. \begin{aligned} \not\models (\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), d_1, \dots, d_l] \approx \\ C_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), d_1, \dots, d_l])). \end{aligned} \right\} \quad (21)$$

Observe that the last relation (21) takes place for any variables  $d_1, \dots, d_l$ . So, for any elements  $\delta_1, \dots, \delta_l \in \mathfrak{A}$ , we have:

$$\left. \begin{aligned} \not\models (\bigvee_{j=1}^k \wedge_{i=1}^m (B_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), \delta_1, \dots, \delta_l] \approx \\ C_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), \delta_1, \dots, \delta_l])). \end{aligned} \right\} \quad (22)$$

Let us consider the following elements of algebra  $\mathfrak{A}$ :

$$\delta_w = \Phi(\tau_{w1}), \quad (23)$$

where  $\tau_{w1} = D_w[\beta_{11}, \dots, \beta_{n1}]$ ,  $w = 1, \dots, l$ . Now, substituting (23) into (22), we also get:

$$\left. \begin{aligned} \not\models (\bigvee_{j=1}^k \wedge_{i=1}^m ( \\ B_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), \Phi(\tau_{11}), \dots, \Phi(\tau_{l1})] \approx \\ C_{ij}[\Phi(\beta_{11}), \dots, \Phi(\beta_{n1}), \Phi(\beta_1), \Phi(\tau_{11}), \dots, \Phi(\tau_{l1})])). \end{aligned} \right\} \quad (24)$$

From this last relation (24), since  $B_{ij}, C_{ij} \in K$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, k$  and according to relations (8) and (9), we have that:

$$\left. \begin{aligned} \not\models \bigvee_{j=1}^k \wedge_{i=1}^m ( \Phi(B_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{l1}]) \approx \\ \Phi(C_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{l1}])). \end{aligned} \right\} \quad (25)$$



This means that for any  $j = 1, \dots, k$ , we have:

$$\left. \begin{aligned} \nexists \bigwedge_{i=1}^m (\Phi(B_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{11}]) \approx \\ \Phi(C_{ij}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{11}])). \end{aligned} \right\} \quad (26)$$

Further it follows that there exists  $i_j, i_j \in \{1, \dots, m\}$ , such that

$$\left. \begin{aligned} \nexists (\Phi(B_{i_j j}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{11}]) \approx \\ \Phi(C_{i_j j}[\beta_{11}, \dots, \beta_{n1}, \beta_1, \tau_{11}, \dots, \tau_{11}])). \end{aligned} \right\}$$

From the last relation it follows that there exist  $r_j, r_j \in \{1, \dots, s\}$ , such that:

$$\left. \begin{aligned} \nexists B_{i_j j}[\beta_{1r_j}, \dots, \beta_{nr_j}, \beta_{r_j}, \tau_{1r_j}, \dots, \tau_{lr_j}] \approx \\ C_{i_j j}[\beta_{1r_j}, \dots, \beta_{nr_j}, \beta_{r_j}, \tau_{1r_j}, \dots, \tau_{lr_j}]. \end{aligned} \right\} \quad (27)$$

The last relation (27) implies also the following:

$$\left. \begin{aligned} \nexists \bigwedge_{i=1}^m B_{ij}[\beta_{1r_j}, \dots, \beta_{nr_j}, \beta_{r_j}, \tau_{1r_j}, \dots, \tau_{lr_j}] \approx \\ C_{ij}[\beta_{1r_j}, \dots, \beta_{nr_j}, \beta_{r_j}, \tau_{1r_j}, \dots, \tau_{lr_j}]. \end{aligned} \right\} \quad (28)$$

Let us remark that the relations (27) and (28) hold for any  $j = 1, \dots, k$ . Therefore, the following relation is true:

$$\left. \begin{aligned} \nexists \bigvee_{j=1}^k \bigwedge_{i=1}^m B_{ij}[\beta_{1r_j}, \dots, \beta_{nr_j}, \beta_{r_j}, \tau_{1r_j}, \dots, \tau_{lr_j}] \approx \\ C_{ij}[\beta_{1r_j}, \dots, \beta_{nr_j}, \beta_{r_j}, \tau_{1r_j}, \dots, \tau_{lr_j}]. \end{aligned} \right\} \quad (29)$$

Comparing relations (29) and (17), we conclude that we get a contradiction.

The theorem is proved.

**Theorem 2.** *Suppose  $\mathfrak{A}$  is an algebra and  $b \in \mathfrak{A}$ . Then the set  $K$  of terms of  $\mathfrak{A}$  that conserve on  $\mathfrak{A}$  the relation  $x = b$  is closed relative to existential expressibility on  $\mathfrak{A}$ .*

The proof of this theorem is almost obvious if we consider in Theorem 1 the mapping  $\Phi(x) = b$ .

## 4 Main results

Consider logics  $L\mathfrak{B}_i$ ,  $i = 1, 2, 3$  of the corresponding algebras  $\mathfrak{B}_i$ , known also as extensions of the logic  $S5$ .

Consider classes of formulas  $\Pi_0, \Pi_1, \Pi_2$ , that conserve on algebra  $\mathfrak{B}_1$  the relations  $x = 0$ ,  $x = 1$ , and  $\neg x = y$ .

**Theorem 3.** *The classes of formulas  $\Pi_0, \Pi_1, \Pi_2$  of the logic  $L\mathfrak{B}_1$  are pre-complete relative to existential expressibility in  $L\mathfrak{B}_1$ .*

According to Theorem 1, these classes are closed as to existential expressibility. By E. Post's results [1],[2], these classes are pre-complete as to expressibility. So, they are also pre-complete as to existential expressibility, too.

Consider elements  $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$  of  $\mathfrak{B}_2$  denoted by  $0, \rho, \sigma, 1$ , respectively. Consider mapping  $f_{10} : \mathfrak{B}_2 \rightarrow \mathfrak{B}_2$  defined by relations:  $f_{10}(x) = 0$ , if  $x \in \{0, \rho\}$  and  $f_{10}(x) = 1$ , if  $x \in \{\sigma, 1\}$ .

Consider classes of formulas  $\Pi_8, \Pi_9, \Pi_{10}$  that conserve on the algebra  $\mathfrak{B}_2$  the relations  $\Box x = y$ ,  $\Diamond x = y$ ,  $f_{10}(x) = y$ . Similar to the previous theorem, we have:

**Theorem 4.** *The classes of formulas  $\Pi_8, \Pi_9, \Pi_{10}$  of the logic  $L\mathfrak{B}_2$  are pre-complete relative to existential expressibility in  $L\mathfrak{B}_2$ .*

The proof is similar to the proof of the previous theorem.

Consider algebra  $\mathfrak{B}_3$ . Denote its elements  $\{(0, 0, 0), (0, 0, 1), (0, 1, 0), (1, 0, 0), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$  by  $\{0, \rho, \mu, \varepsilon, \sigma, \nu, \omega, 1\}$ .

Consider mappings  $f_2, f_3, f_4 : \mathfrak{B}_3 \rightarrow \mathfrak{B}_3$  defined in tabular form as in Table 1 below (see [10, p. 168]).

$p$	0	$\rho$	$\mu$	$\varepsilon$	$\omega$	$\nu$	$\sigma$	1
$f_2$	0	$\sigma$	$\sigma$	$\sigma$	$\rho$	$\rho$	$\rho$	1
$f_3$	0	$\rho$	$\nu$	$\omega$	$\varepsilon$	$\mu$	$\sigma$	1
$f_4$	0	$\sigma$	$\omega$	$\nu$	$\mu$	$\varepsilon$	$\rho$	1

Table 1. Functions on  $\mathfrak{B}_3$  [10, p. 168]

Consider [10] classes of formulas  $\Pi_{21}, \Pi_{22}, \Pi_{23}$  that conserve on the algebra  $\mathfrak{B}_3$  the relations  $f_2(x) = y, f_3(x) = y, f_4(x) = y$ .

**Theorem 5.** *The classes of formulas  $\Pi_{21}, \Pi_{22}, \Pi_{23}$  of the logic  $L\mathfrak{B}_3$  are pre-complete relative to existential expressibility in  $L\mathfrak{B}_3$ .*

**Remark 1.** *Only the class formulas  $\Pi_2$  from the above-mentioned classes contain the para-consistent negation. So, if a system  $\Sigma$  of formulas containing para-consistent negation is complete as to existential expressibility in  $S5$ , it should satisfy the relation:  $\Sigma \not\subset \Pi_2$ .*

Now we can give necessary and sufficient conditions for a system of boolean functions to be complete relative to existential expressibility mentioned in [8].

**Theorem 6.** *Consider the Boolean algebra  $\mathfrak{B}_1 = (\{0, 1\}; \&, \vee, \neg, 0, 1)$ . The system  $\Sigma$  of boolean functions is complete relative existential expressibility if and only if it conserves none of the relations on  $\mathfrak{B}_1$ :  $x = 0, x = 1, \text{ and } x = \neg y$ .*

**Proof.** Each class of functions that conserve the corresponding relation is closed (according to Theorem 1) as to existential expressibility. It is also known that these classes are distinct [1], and according to [3], the constants 0 and 1 are expressible via  $\Sigma$ . By force of the Example 1, we conclude the system  $\Sigma$  is complete as to existential expressibility.

## 5 Conclusions

Conditions for a system of formulas containing para-consistent negations to be existential expressible in the logic  $S5$  are only necessary conditions.

The discovery of all necessary and sufficient conditions for a system of formulas of  $S5$  to be complete as to existential expressibility may follow the following procedure:

- Consider possible classes of formulas as possible candidates that comply with the conditions stated in Theorem 1.

- Apply the principle from simple to complex, i.e., start with the corresponding 2-, 4-, 8-valued algebras.
- Examine initially classes defined by 1-valued functions  $\Phi$ .
- As the dimension of the algebraic model of the logic under consideration may increase (2-valued, 4-valued, 8-valued, 16-valued, 32-valued), it is useful to filter the possible  $\Phi$  functions mentioned above. A relatively simple way is to consider different  $\Phi$  functions and examine the relation of the formulas of the logic on the corresponding algebraic model relative to classes defined by those  $\Phi$  functions. This will allow establishing a possible inclusion of the classes defined by  $\Phi$  functions in each other. So, a software, for example, something similar to <http://tinyurl.com/4ut3f7em>, after some adaptation, may help in filtering unnecessary classes of formulas. The above theorems are useful to assure the closure of the corresponding classes relative to existential expressibility.

This paper is the extended and revised version of the conference paper [14] presented at WIIS 2023.

**Acknowledgments.** National Agency for Research and Development has supported part of the research for this paper through the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

## References

- [1] E. Post, "Introduction to a general theory of elementary propositions," *Amer. J. Math.*, vol. 43, pp. 163–185, 1921.
- [2] E. L. Post, *The Two-Valued Iterative Systems of Mathematical Logic.(AM-5), Volume 5*. Princeton University Press, 2016, vol. 5.
- [3] S. V. Jablonskij, *Introduction to discrete mathematics*. Moscow: Nauka, 1986.

- [4] A. V. Kuznetsov, "On problems of identity and criteria of functional completeness," in *Proceedings of the 3rd Allunion Congress of Mathematics*, Moscow, Ed., vol. 2, 1956, pp. 145–146.
- [5] I. G. Rosenberg, "La structure des fonctions de plusieurs variables sur un ensemble fini," *CR Acad. Sci. Paris*, vol. 260, pp. 3817–3819, 1965.
- [6] J.-Y. Beziau, "S5 is a paraconsistent logic and so is first-order classical logic." *Logical Investigations*, vol. 9, pp. 301–309, Oct. 2002. [Online]. Available: <https://logicalinvestigations.ru/article/view/213>
- [7] S. J. Scroggs, "Extensions of the lewis system s51," *The Journal of Symbolic Logic*, vol. 16, no. 2, pp. 112–120, 1951.
- [8] A. V. Kuznetsov, "On tools for detection of non-deduction or non-expressibility," in *Logical deduction*, M. Nauka, Ed. Nauka, 1979, pp. 5–33.
- [9] A. V. Kuznetsov, "On functional expressibility in the superintuitionistic logics," *Matematicheskie issledovaniya*, vol. 6, no. 4, pp. 75–122, 1971. [Online]. Available: <http://eudml.org/doc/189110>
- [10] M. Rața, *Expressibility in propositional calculi*. Chișinău: Știința, 1991.
- [11] M. F. Ratsa, "On functional expressibility in intuitionistic propositional logic," in *Problems of cybernetics*. Moscow: Nauka, 1982, vol. 39, pp. 107–150.
- [12] J. Robinson, "Existential definability in arithmetic," *Transactions of the American Mathematical Society*, vol. 72, no. 3, pp. 437–449, 1952.
- [13] S. Burris and H. P. Sankappanavar, *A course in universal algebra*. Springer, 1981, vol. 78.
- [14] A. Rusu and E. Rusu, "On some pre-complete relative to positive expressibility classes of formulas in the 8-valued para-consistent

extension of the logic  $s_5$ ,” in *Proceedings of Workshop on Intelligent Information Systems: WIIS2023*, 2023, pp. 199–206.

Andrei Rusu<sup>1,3</sup>, Elena Rusu<sup>2</sup>

Received September 30, 2023

Accepted December 13, 2023

<sup>1</sup>Vladimir Andrunavhievici Institute of Mathematics  
and Computer Science, State University of Moldova  
5, Academiei street, Chişinău, Republic of Moldova, MD2028  
ORCID: <https://orcid.org/0000-0002-0259-3060>  
E-mail: [andrei.rusu@math.md](mailto:andrei.rusu@math.md)

<sup>2</sup>Dep. of Mathematics, Technical University of Moldova  
168, Stefan cel Mare bd, Chisinau, Republic of Moldova, MD-2004  
ORCID: <https://orcid.org/0000-0002-2473-0353>  
E-mail: [elena.rusu@mate.utm.md](mailto:elena.rusu@mate.utm.md)

<sup>3</sup>Dep. of Mathematics and Informatics, Ovidius University of Constanţa  
124, Mamaia Bd., Constanţa, Romania, 900527  
E-mail: [andrei.rusu@365.univ-ovidius.ro](mailto:andrei.rusu@365.univ-ovidius.ro)